Chapter 1: Extract, Transform, and Load

Logical Operators in the R Environment			
<	Less than	>	Greater than
<=	Less than or equal to	>=	Greater than or equal to
==	Equal to	!=	Not equal to
%in%	Group membership	&, |, !, xor, any, and all	Boolean operators
is.na	Is NA	!is.na	Is not NA

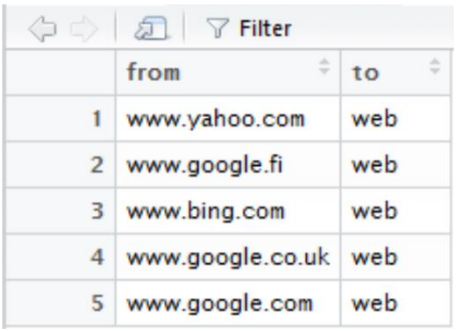
	season	casual
1	1	2
2	1	1

	season	casual	revenue
1	1	2	10
2	1	1	5
3	1	4	20
4	1	1	5
5	1	1	5
6	1	1	5
7	1	1	5
8	1	3	15
9	2	3	15
10	2	1	5

Filter

	season	sum(casual)	sum(revenue)
1	1	14	70
2	2	4	20

Chapter 2: Data Cleaning

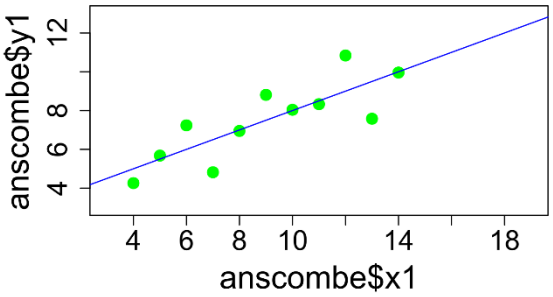


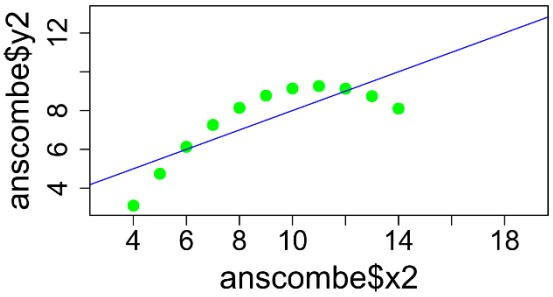
Filter

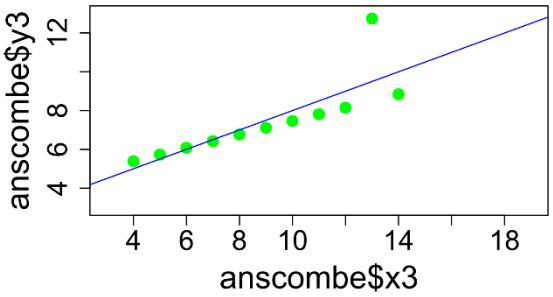
	from	to
1	www.yahoo.com	web
2	www.google.fi	web
3	www.bing.com	web
4	www.google.co.uk	web
5	www.google.com	web

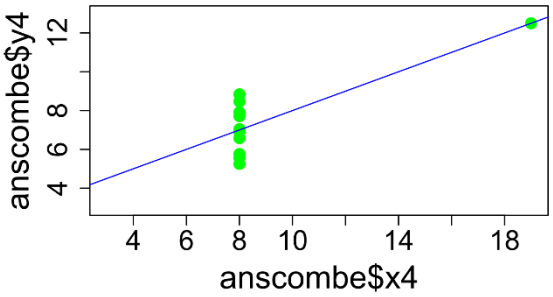
Chapter 3: Exploratory Data Analysis

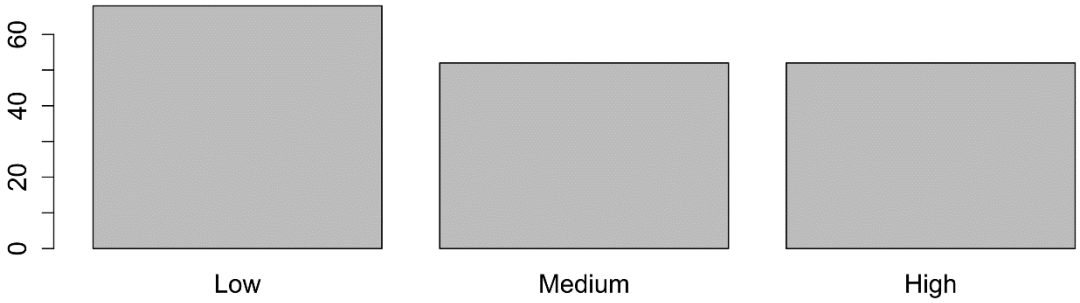
Question → Models → Answer

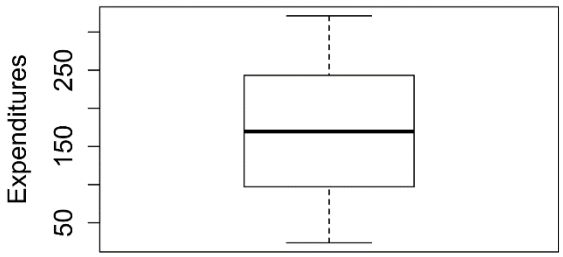


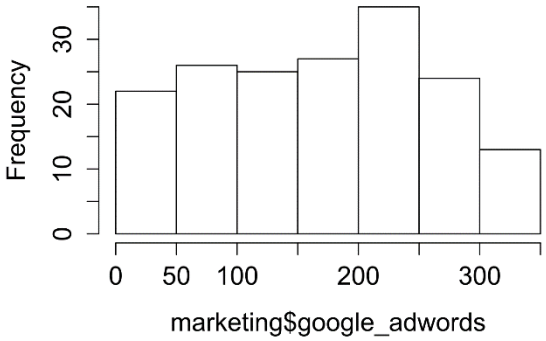


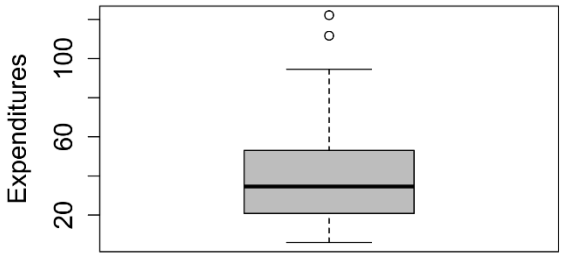


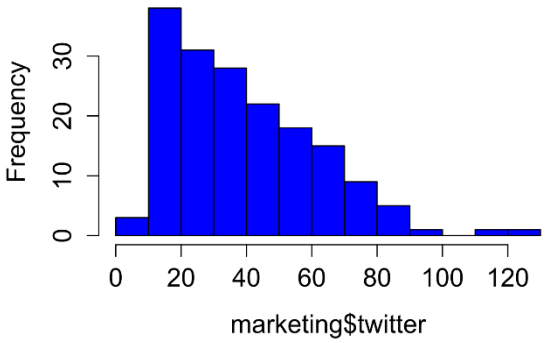


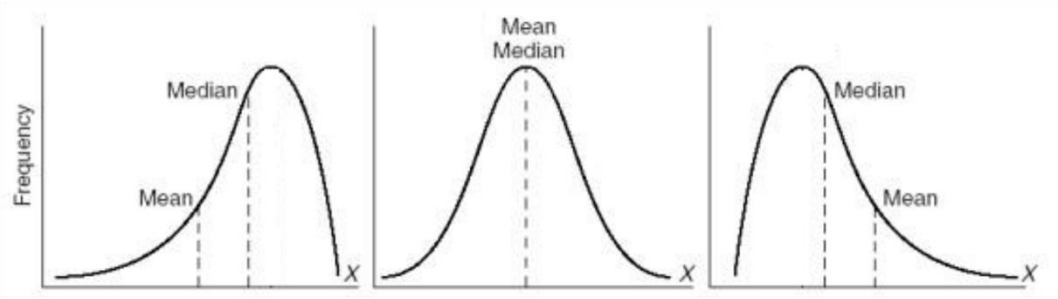




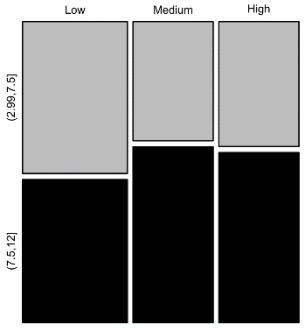




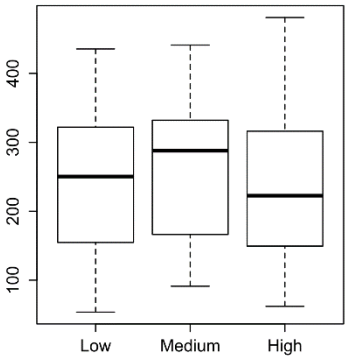




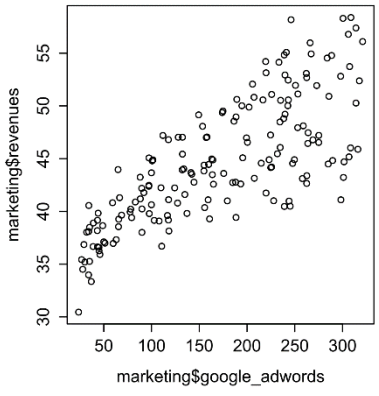
Factor / Factor

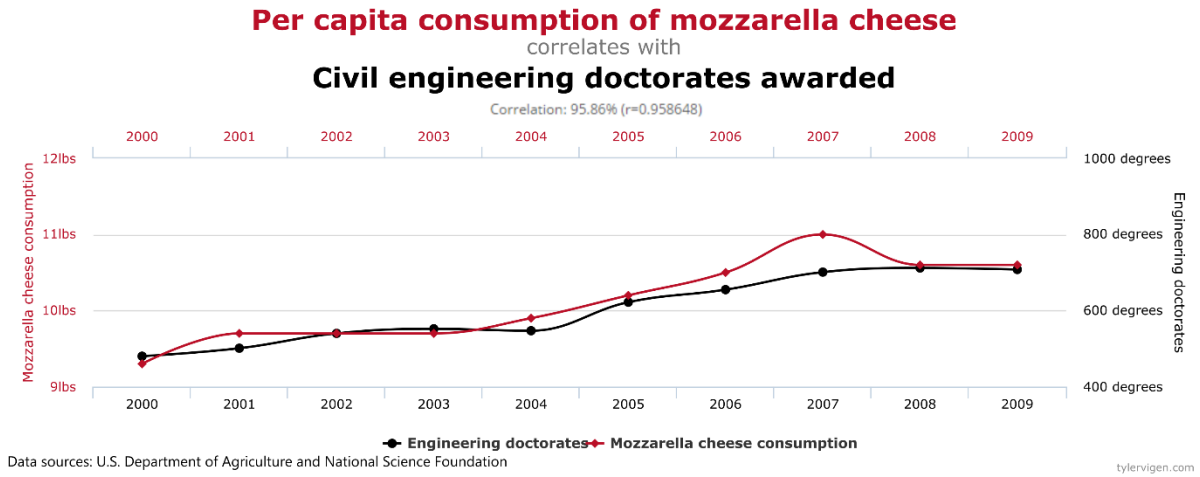


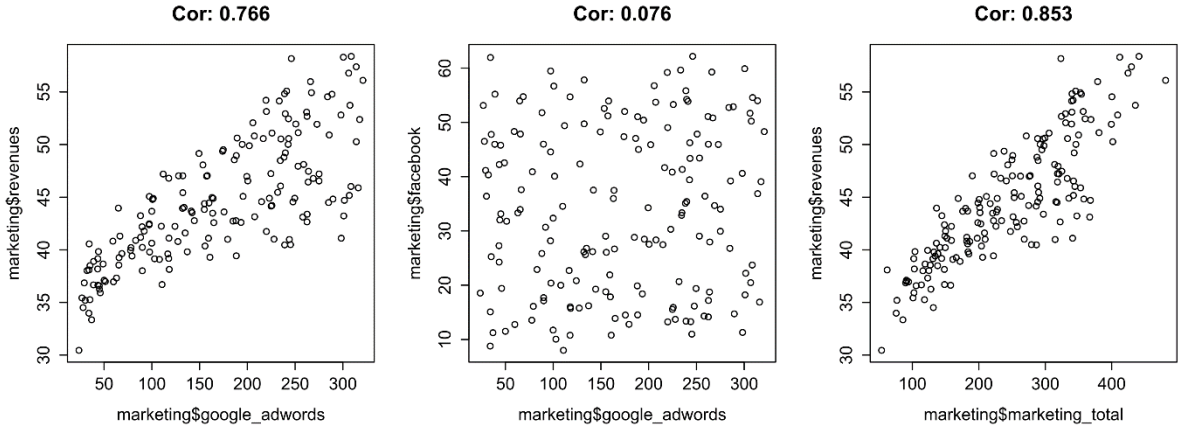
Factor / Numeric

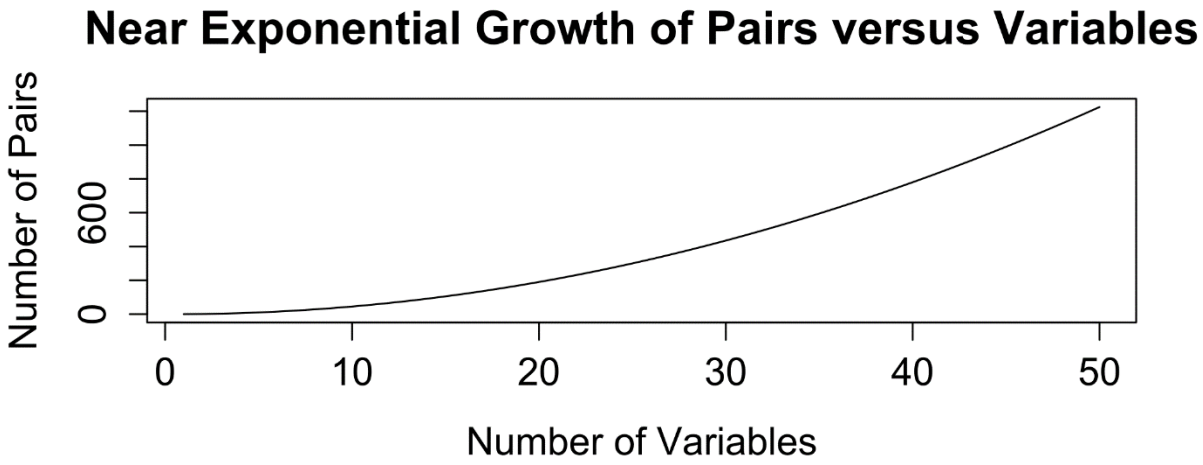


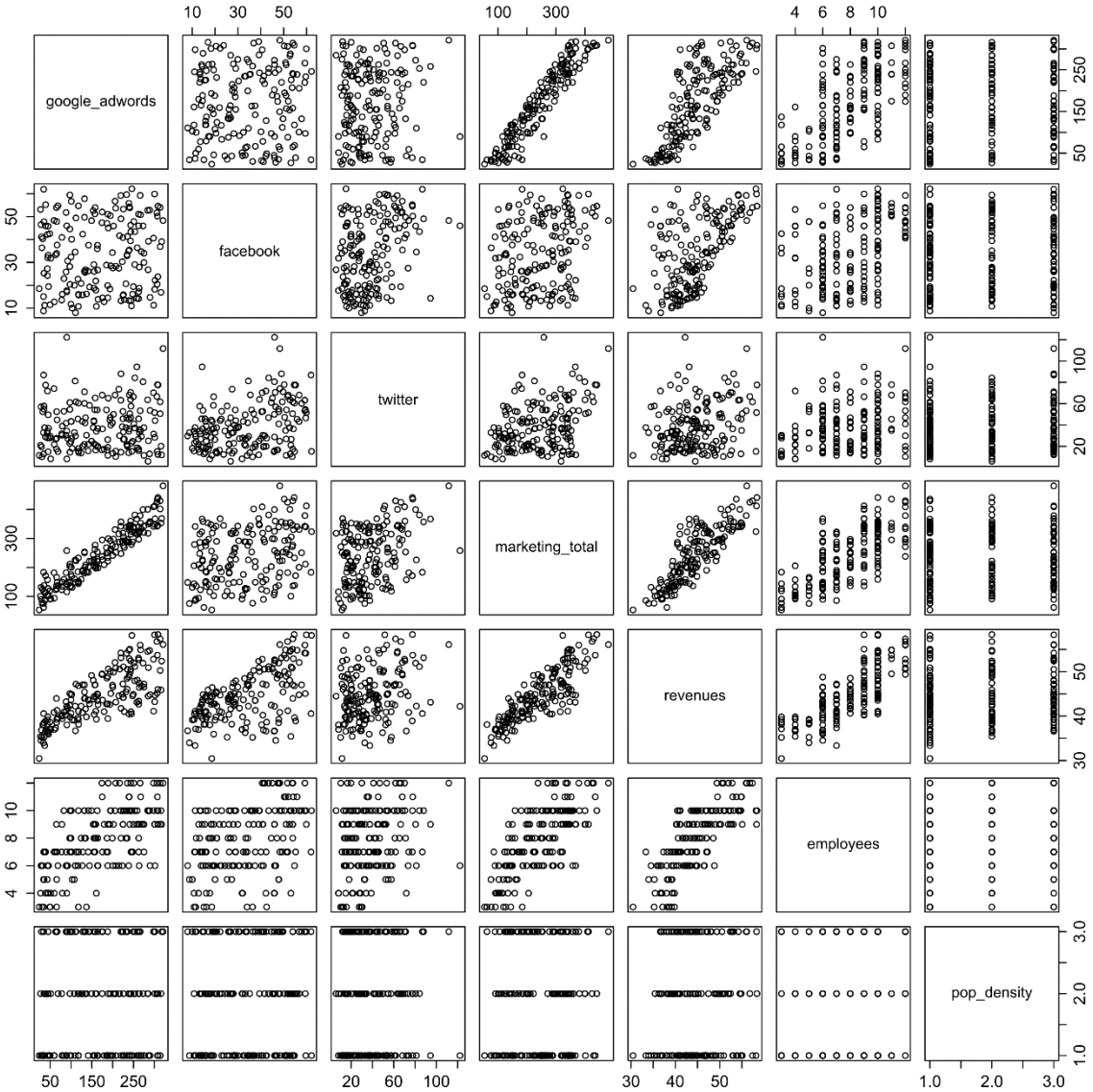
Numeric / Numeric




Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded
Correlation: 95.86% (r=0.958648)
Mozzarella cheese consumption
Engineering doctorates
Engineering doctorates
Mozzarella cheese consumption
Data sources: U.S. Department of Agriculture and National Science Foundation
tylervigen.com


Cor: 0.766
Cor: 0.076
Cor: 0.853
marketing$revenues
marketing$google_adwords
marketing$facebook
marketing$google_adwords
marketing$revenues
marketing$marketing_total

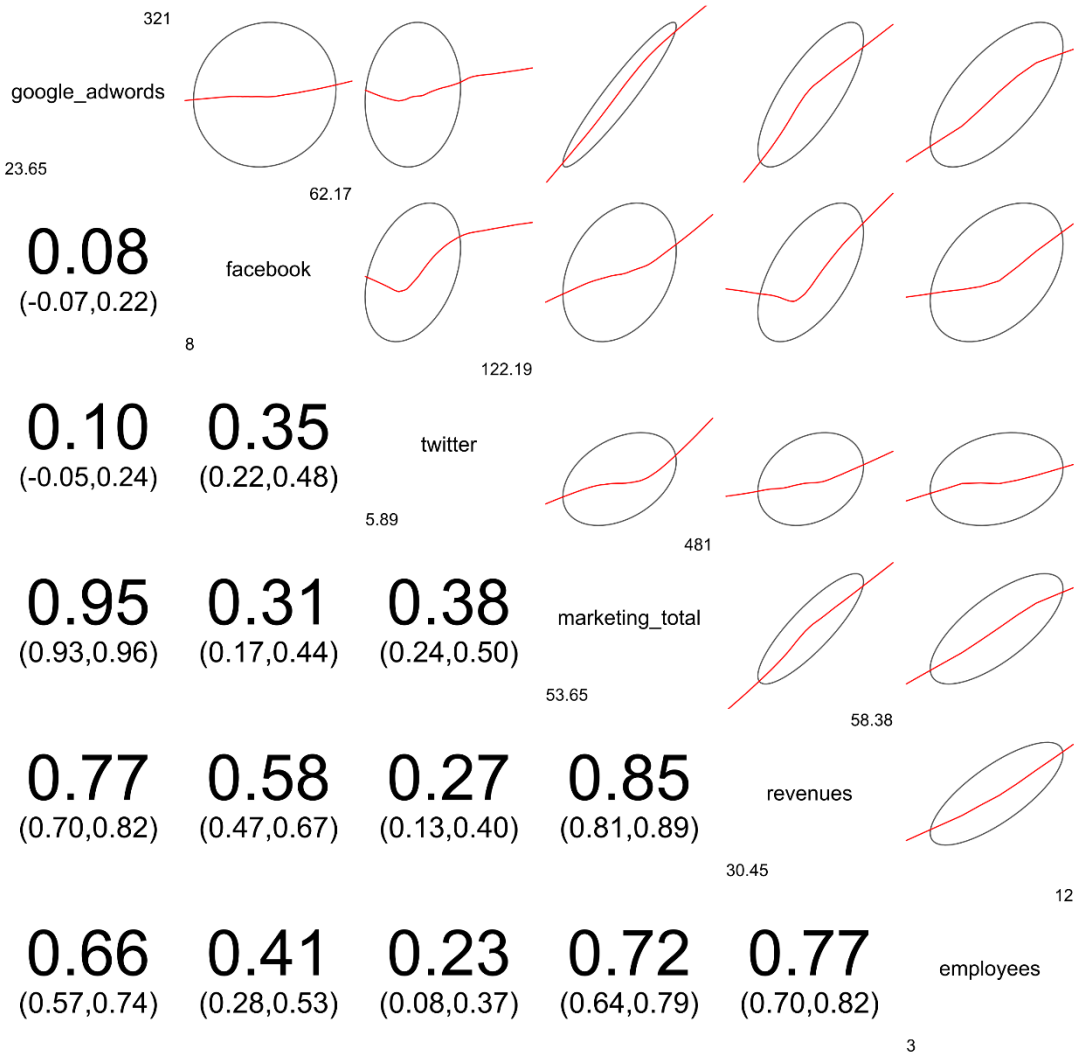

Near Exponential Growth of Pairs versus Variables
Number of Pairs
Number of Variables



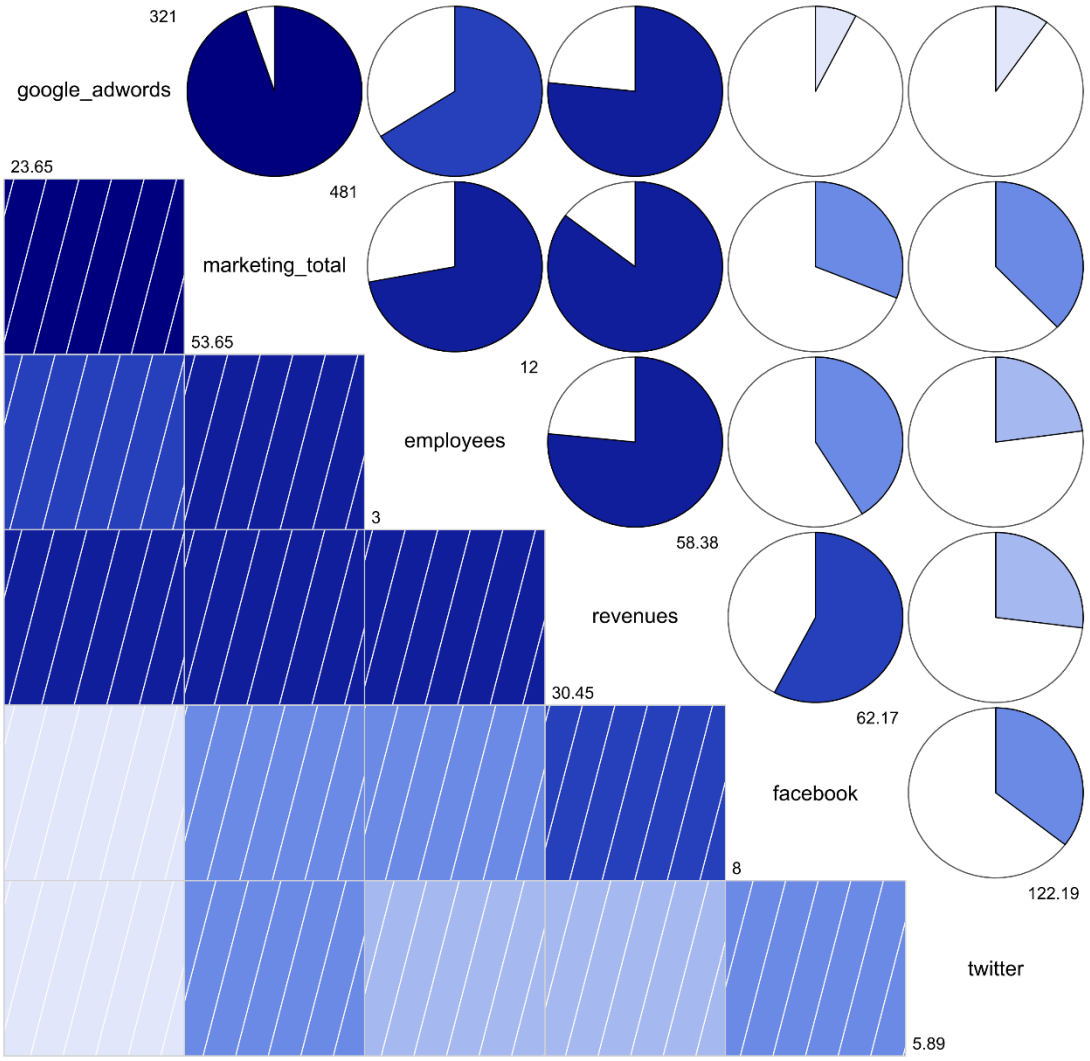
	google_adwords	facebook	twitter	marketing_total	revenues	employees
google_adwords	1.00000000	0.07643216	0.0989750	0.9473566	0.7662461	0.6610312
facebook	0.07643216	1.00000000	0.3543410	0.3102232	0.5778213	0.4101966
twitter	0.09897500	0.35434096	1.0000000	0.3758691	0.2696854	0.2290618
marketing_total	0.94735659	0.31022316	0.3758691	1.0000000	0.8530354	0.7210171
revenues	0.76624608	0.57782131	0.2696854	0.8530354	1.0000000	0.7656857
employees	0.66103123	0.41019661	0.2290618	0.7210171	0.7656857	1.0000000

```
Call:corr.test(x = marketing[, 1:6])
Correlation matrix
                google_adwords facebook twitter marketing_total revenues employees
google_adwords            1.00     0.08    0.10            0.95     0.77      0.66
facebook                  0.08     1.00    0.35            0.31     0.58      0.41
twitter                   0.10     0.35    1.00            0.38     0.27      0.23
marketing_total           0.95     0.31    0.38            1.00     0.85      0.72
revenues                  0.77     0.58    0.27            0.85     1.00      0.77
employees                 0.66     0.41    0.23            0.72     0.77      1.00
Sample Size
[1] 172
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                google_adwords facebook twitter marketing_total revenues employees
google_adwords            0.00     0.39    0.39               0        0      0.00
facebook                  0.32     0.00    0.00               0        0      0.00
twitter                   0.20     0.00    0.00               0        0      0.01
marketing_total           0.00     0.00    0.00               0        0      0.00
revenues                  0.00     0.00    0.00               0        0      0.00
employees                 0.00     0.00    0.00               0        0      0.00
```

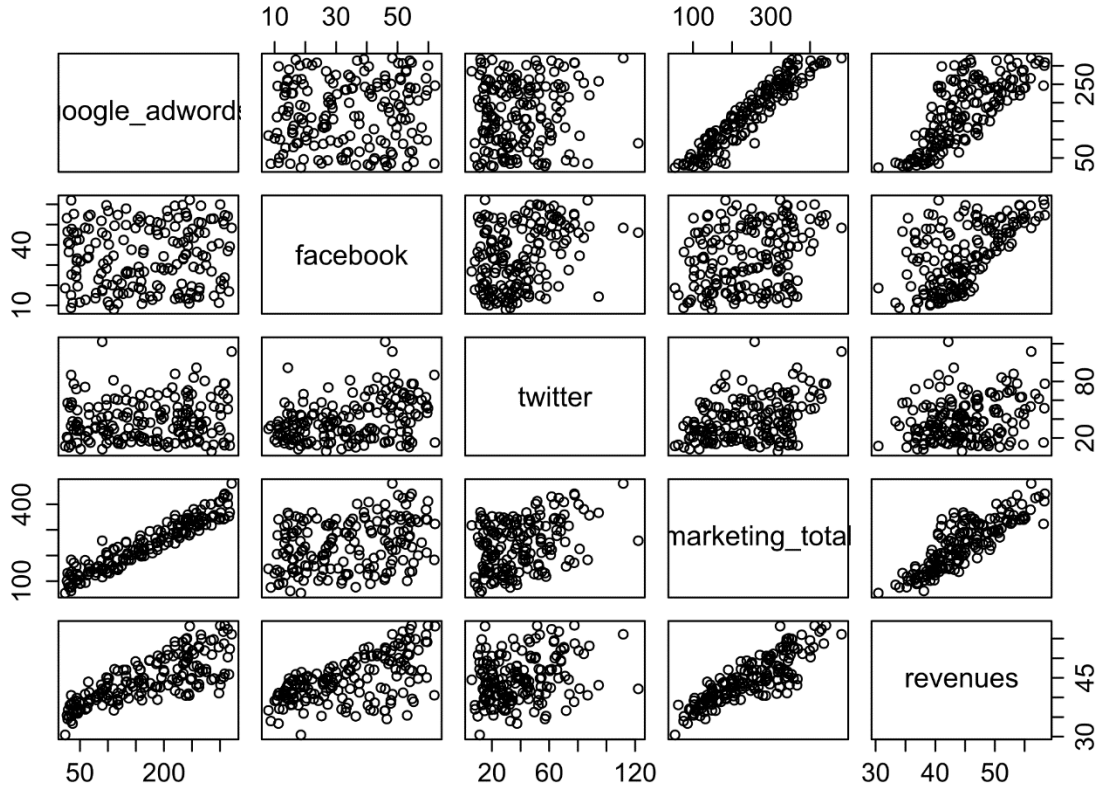
Correlogram of Marketing Data, Unordered

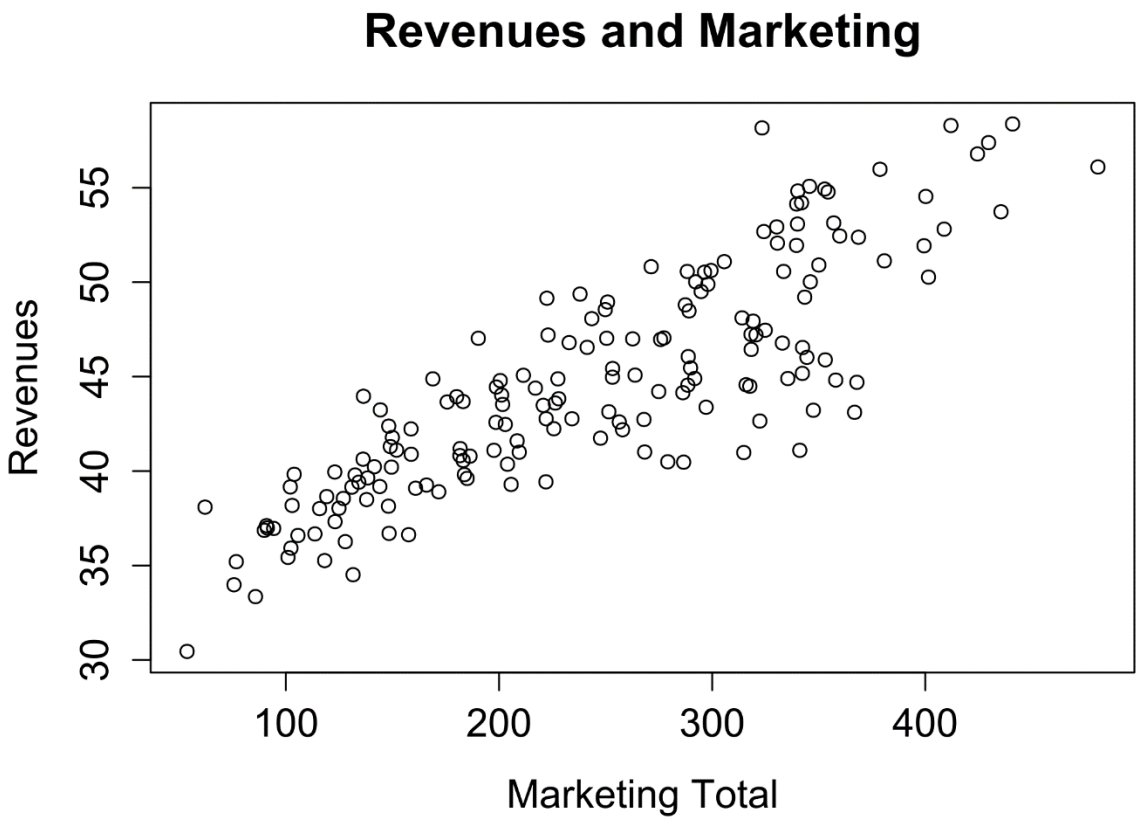


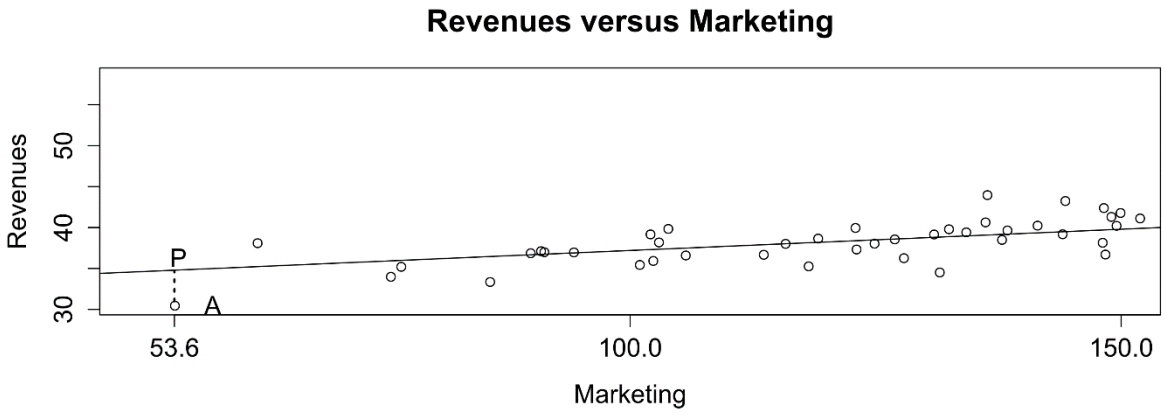
Correlogram of Marketing Data, Ordered

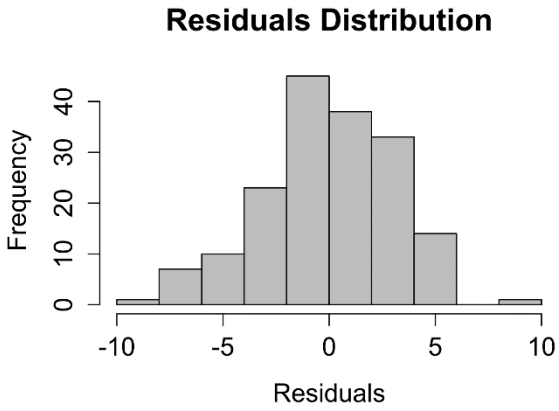


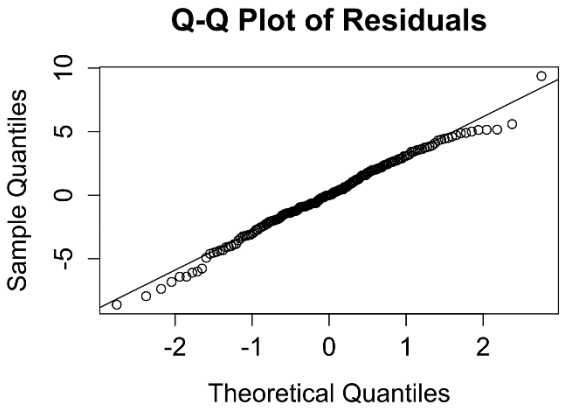
Chapter 4: Linear Regression for Business

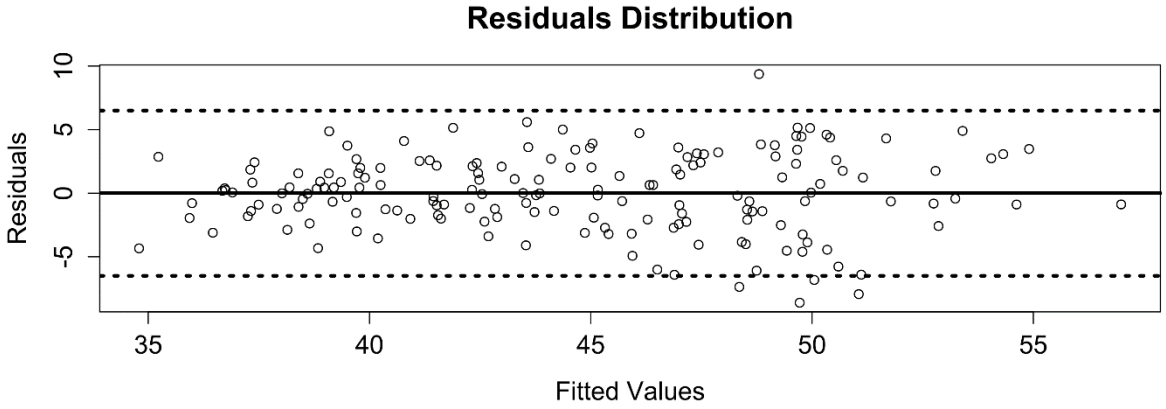


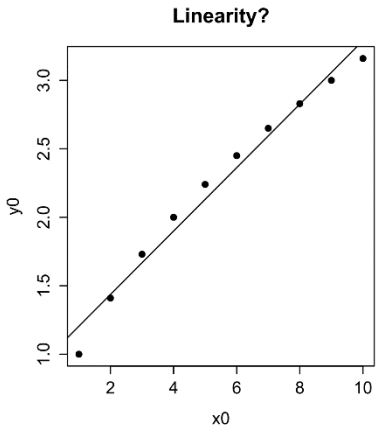

Revenues and Marketing
Revenues
30
35
40
45
50
55
100
200
300
400
Marketing Total

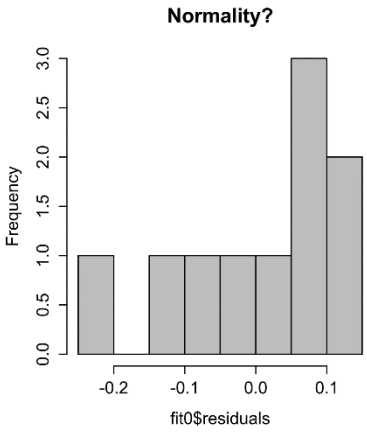

Revenues versus Marketing
Revenues
30
40
50
P
A
53.6
100.0
150.0
Marketing

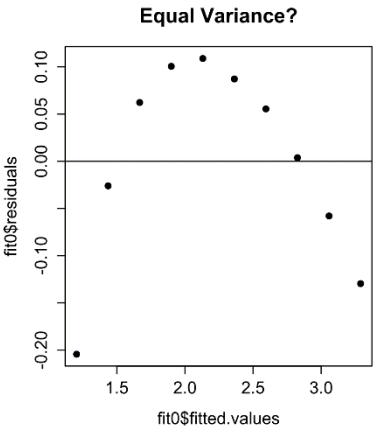

Residuals Distribution
Frequency
0
10
20
30
40
-10
-5
0
5
10
Residuals

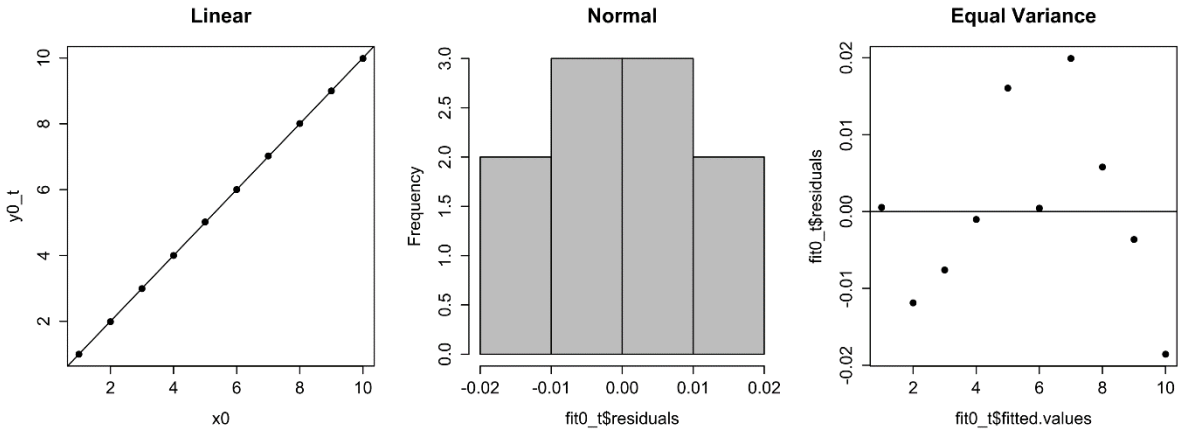

Q-Q Plot of Residuals
Sample Quantiles
-5
0
5
10
-2
-1
0
1
2
Theoretical Quantiles

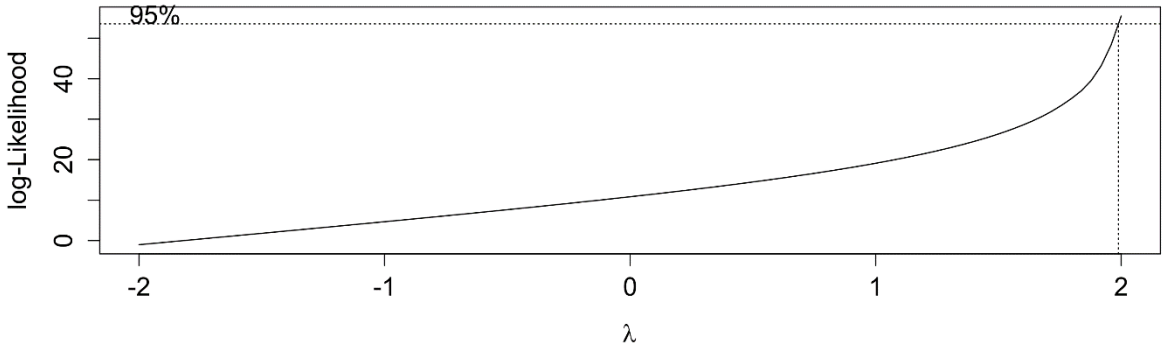

Residuals Distribution
Residuals
-5
0
5
10
35
40
45
50
55
Fitted Values

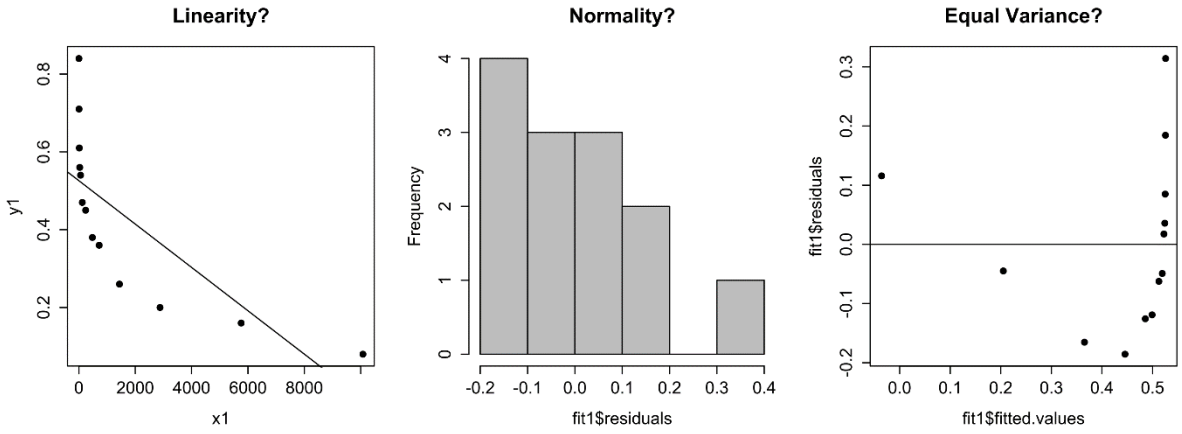

Linearity?
y0
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
x0

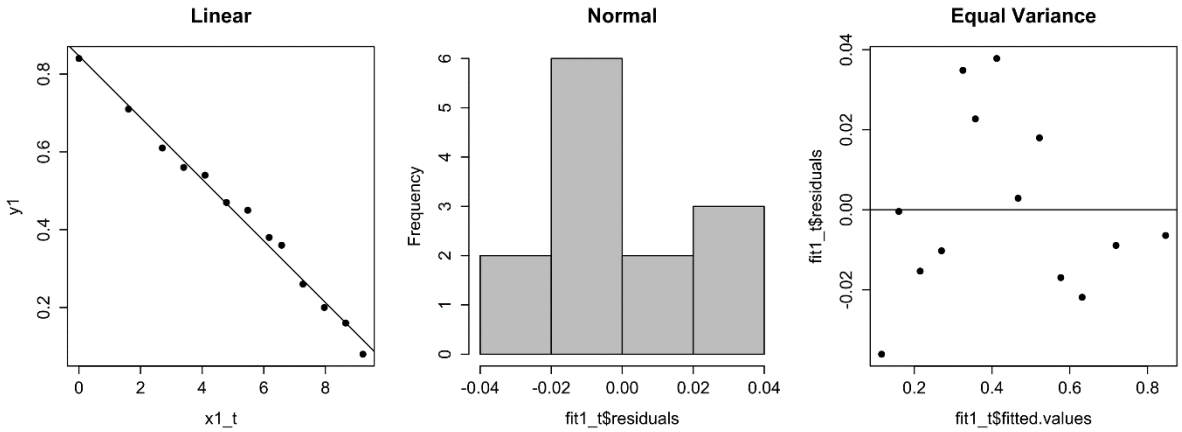

Normality?
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-0.2
-0.1
0.0
0.1
fit0$residuals

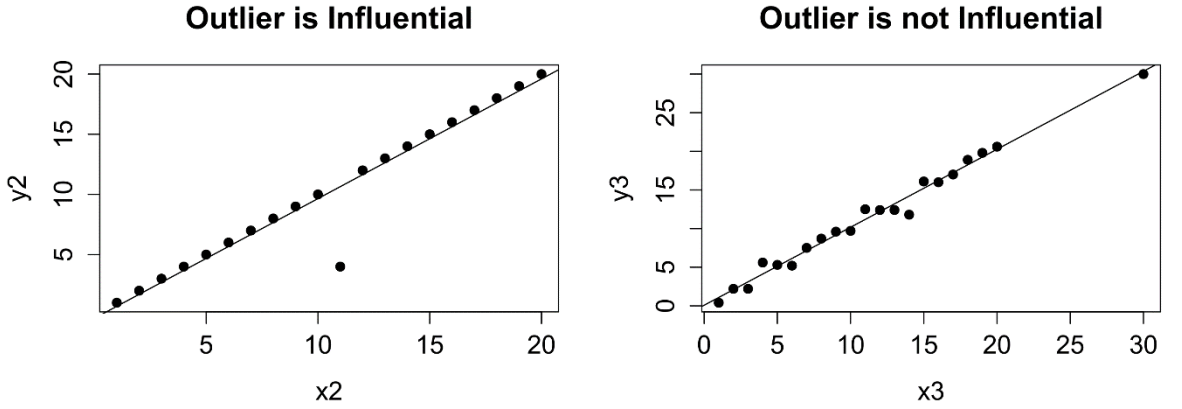

Equal Variance?
fit0$residuals
-0.20
-0.10
0.00
0.05
0.10
1.5
2.0
2.5
3.0
fit0$fitted.values

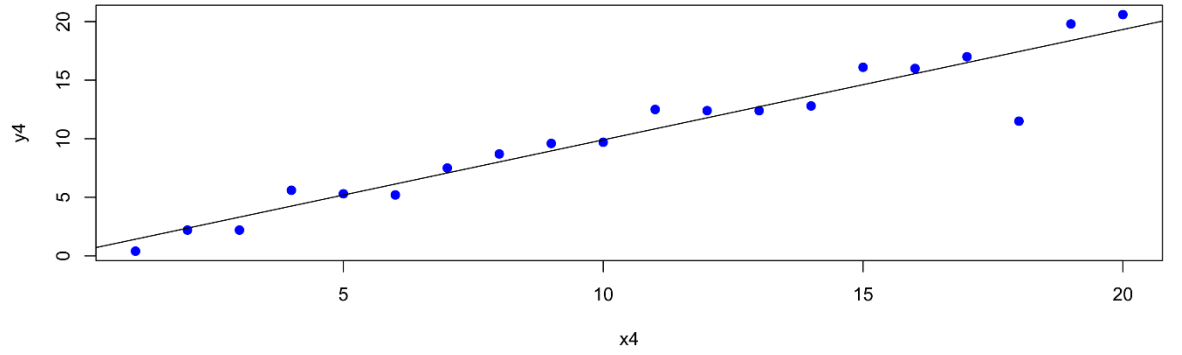

Linear
y0_t
x0
Normal
Frequency
fit0_t$residuals
Equal Variance
fit0_t$residuals
fit0_t$fitted.values

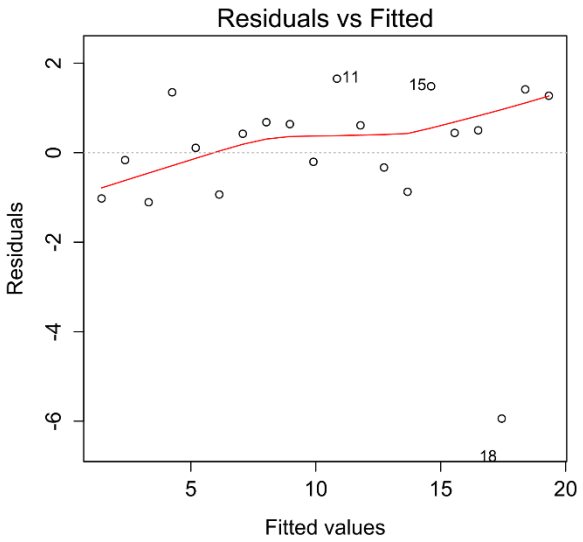

95%
log-Likelihood
λ

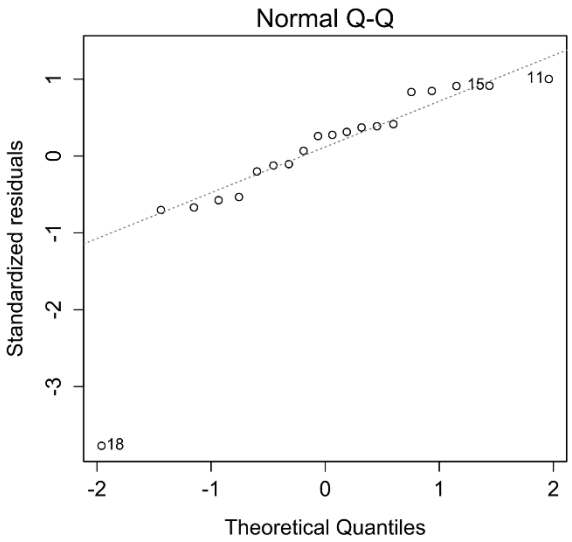

Linearity?
y1
x1
Normality?
Frequency
fit1$residuals
Equal Variance?
fit1$residuals
fit1$fitted.values

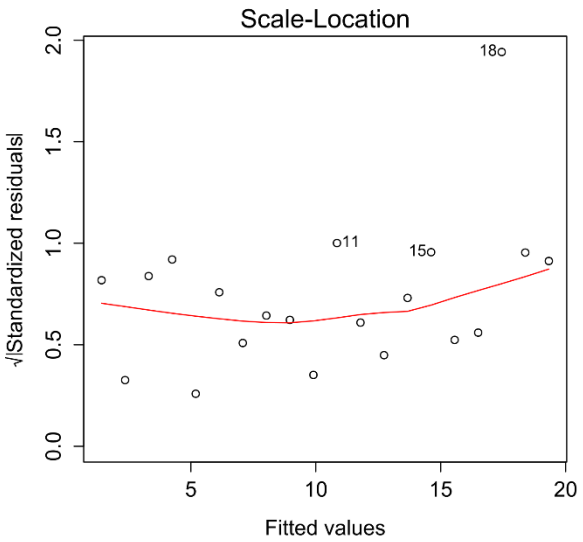

Linear
y1
x1_t
Normal
Frequency
fit1_t$residuals
Equal Variance
fit1_t$residuals
fit1_t$fitted.values

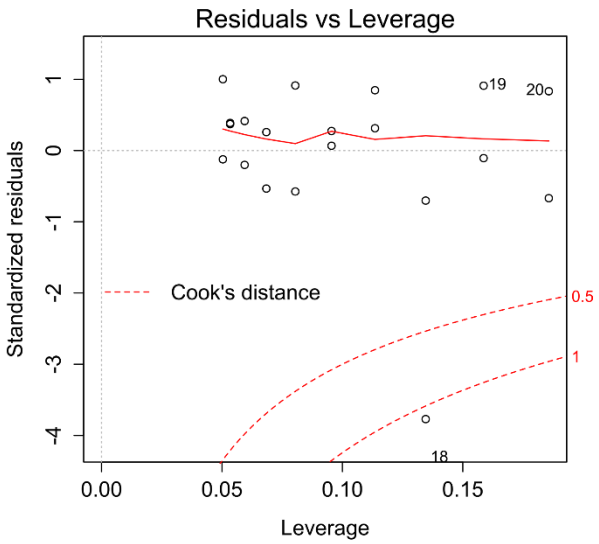

Outlier is Influential
y2
x2
Outlier is not Influential
y3
x3

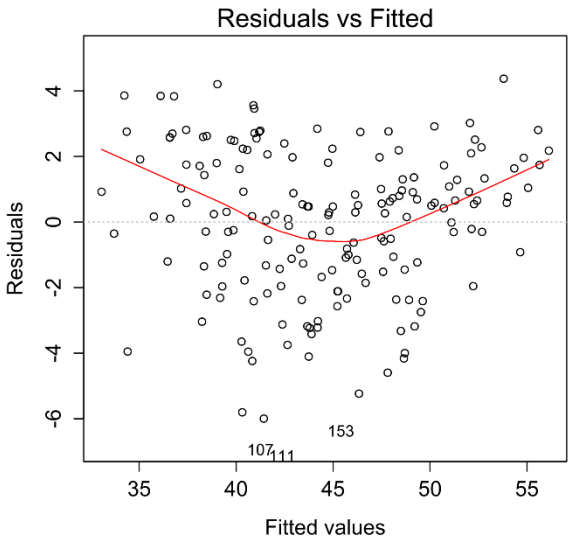

y4
x4

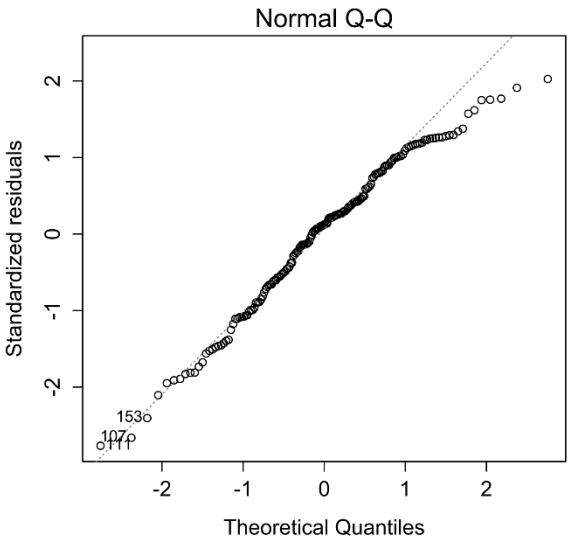

Residuals vs Fitted
Residuals
Fitted values
11
15
18

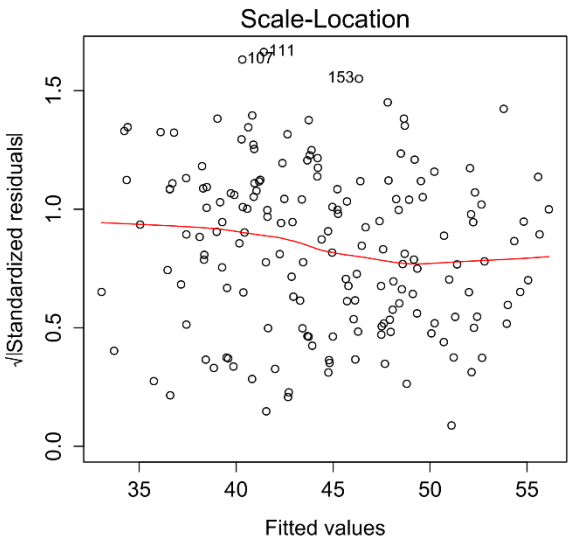

Normal Q-Q
Standardized residuals
Theoretical Quantiles
15
11
18

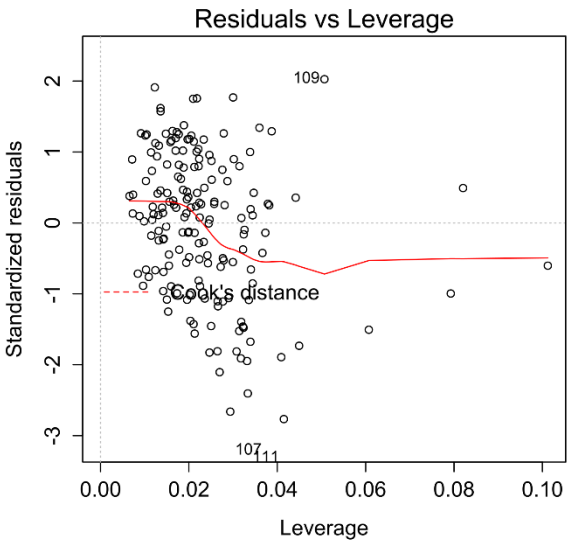

Scale-Location
√|Standardized residuals|
Fitted values
18
11
15


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
1
19
20
18

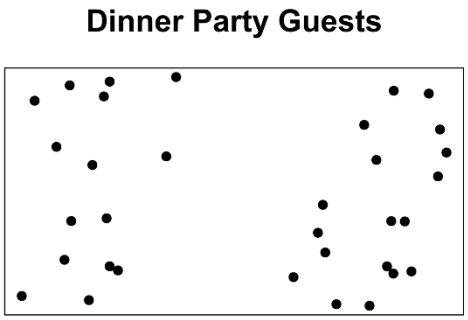

Residuals vs Fitted
Residuals
Fitted values
153
107
111

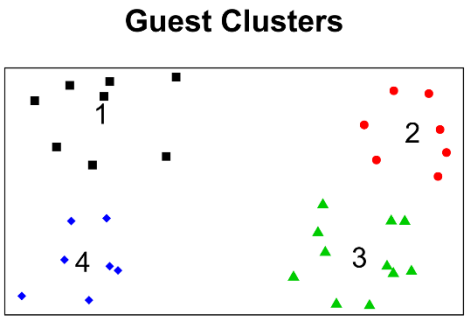

Normal Q-Q
Standardized residuals
Theoretical Quantiles
153
107
111

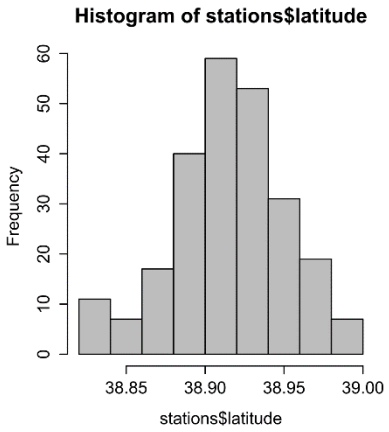

Scale-Location
√|Standardized residuals|
Fitted values
111
107
153

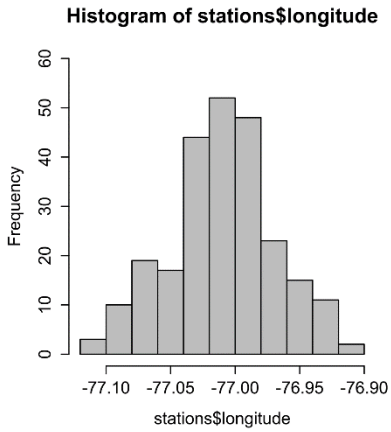

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
109
107
111

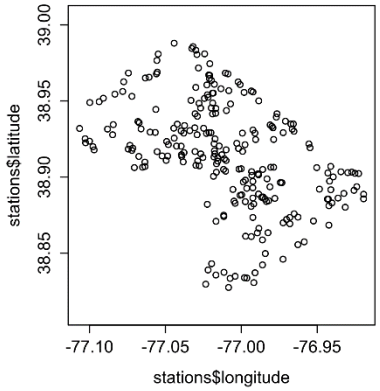
Chapter 5: Data Mining with Cluster Analysis



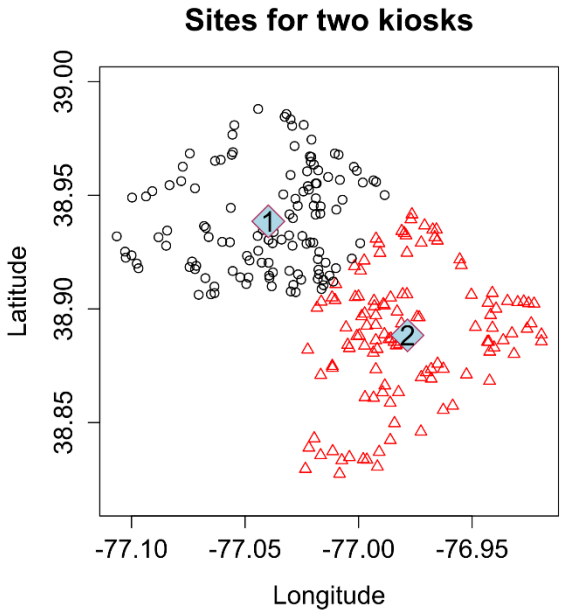


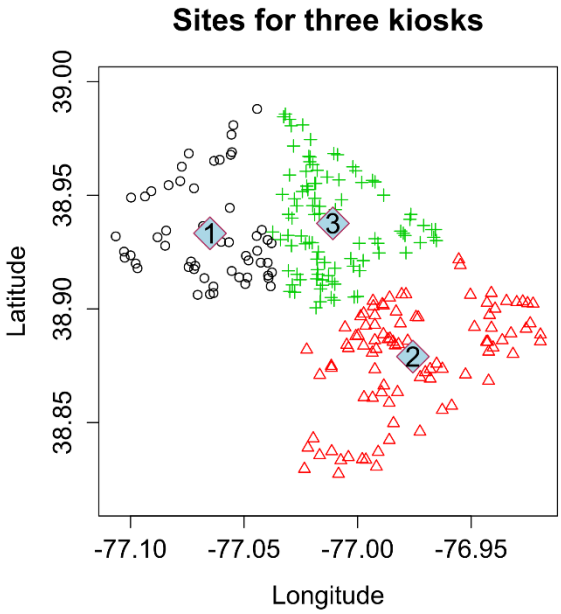


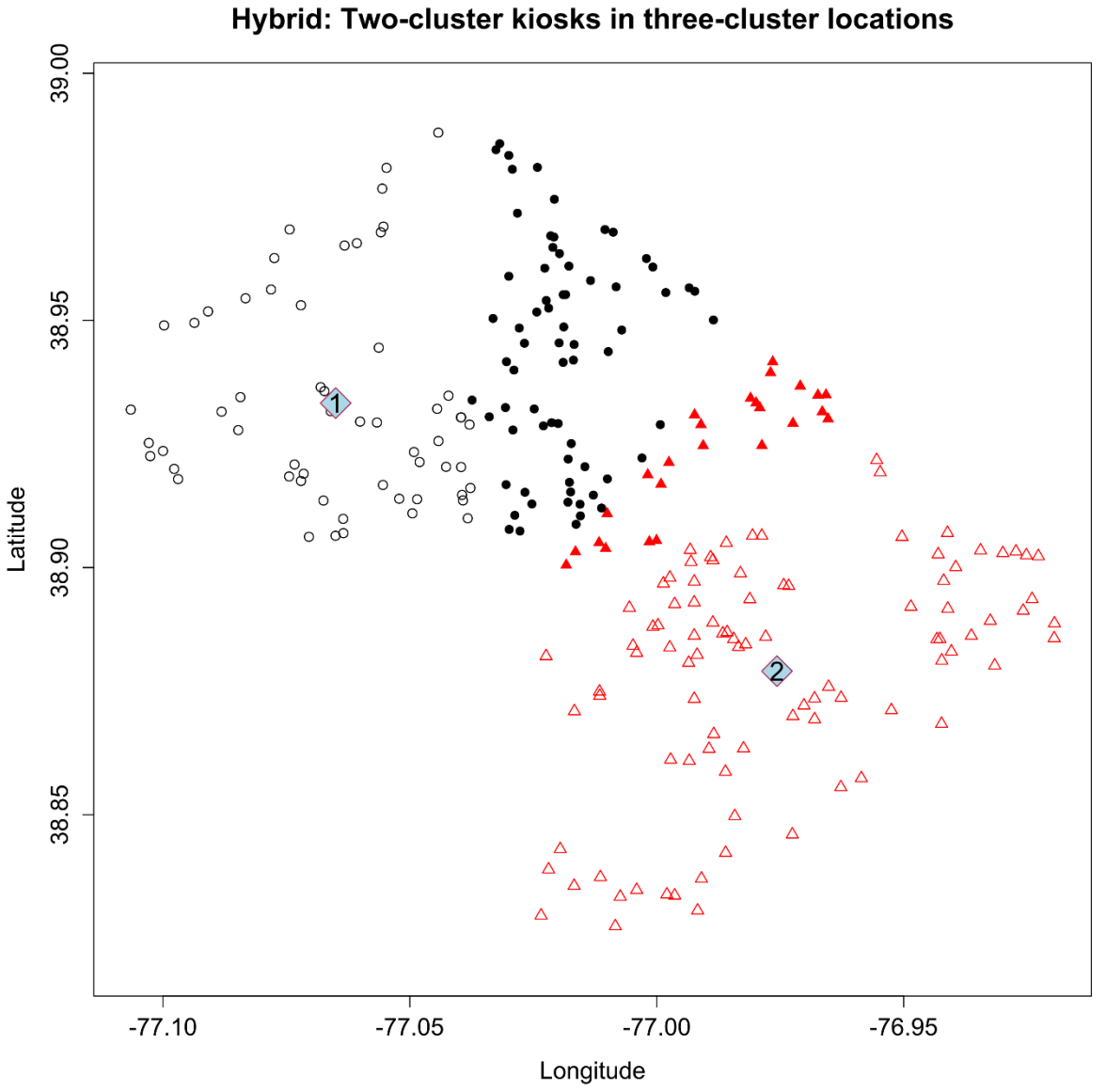


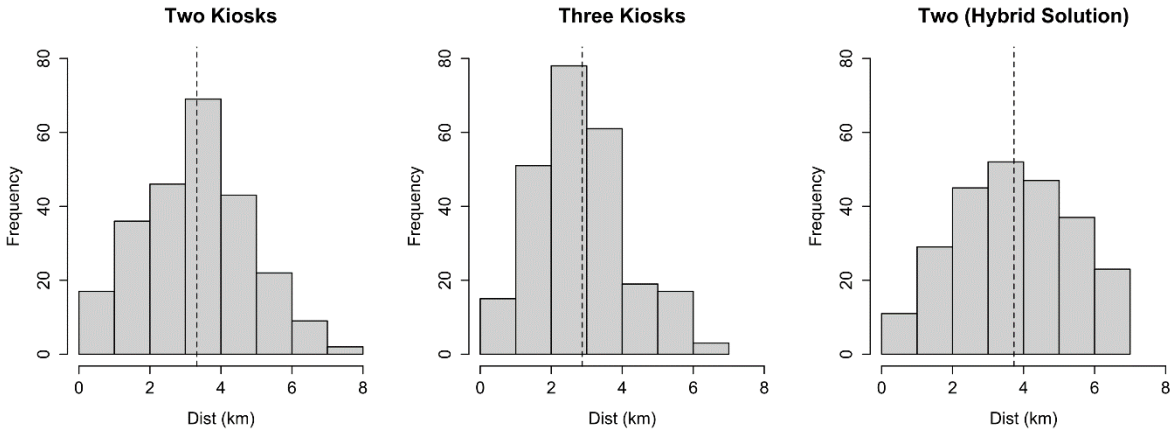


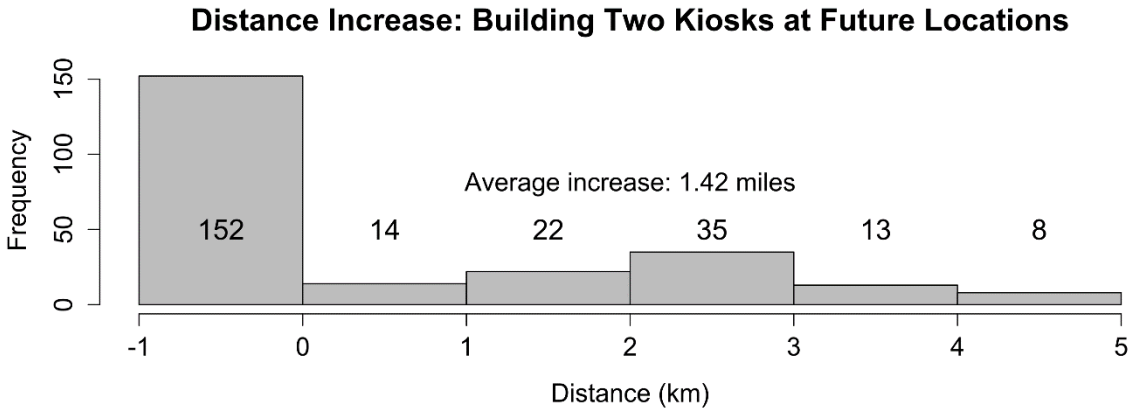

```
K-means clustering with 3 clusters of sizes 57, 93, 94

Cluster means:
  latitude longitude
1 38.93327 -77.06502
2 38.87904 -76.97566
3 38.93765 -77.01089

Clustering vector:
  [1] 3 3 1 2 1 1 2 2 2 2 1 2 2 2 3 1 1 3 1 2 2 3 3 2 1 3 1 3 3 2 2 2 3 1 3 1 3 2
 [39] 3 3 2 3 1 3 2 3 2 1 2 1 2 3 2 2 2 2 2 2 3 2 1 1 1 3 2 2 2 3 3 3 3 3 3 3 2 3
 [77] 3 3 2 1 2 1 1 1 3 1 2 3 2 2 3 2 1 2 2 3 1 2 3 1 3 2 1 1 3 3 1 3 3 3 3 3 2 3
[115] 1 2 3 3 2 3 2 3 1 1 1 2 2 2 2 2 3 3 1 1 3 2 2 2 1 3 3 3 3 1 3 1 1 3 3 2 1 3
[153] 2 3 3 2 1 2 3 3 2 2 2 2 3 2 1 1 3 2 3 3 3 2 1 3 2 3 3 1 2 2 2 1 2 2 2 1 2 3
[191] 2 2 1 3 2 2 2 3 3 3 1 3 1 1 3 2 3 1 3 3 3 2 2 3 3 2 1 3 2 3 3 1 3 2 1 2 2 3
[229] 2 2 3 3 3 1 2 2 2 1 2 1 2 3 2 3

within cluster sum of squares by cluster:
[1] 0.04715762 0.12261951 0.07588127
 (between_SS / total_SS =  65.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

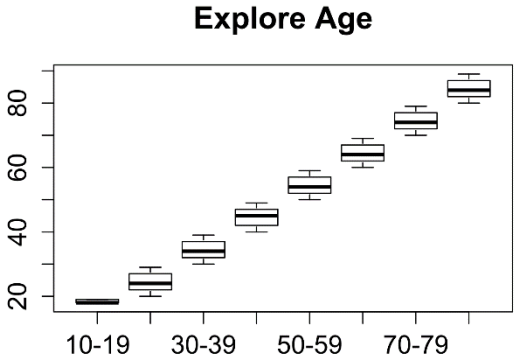

Sites for two kiosks
Latitude
Longitude
39.00
38.95
38.90
38.85
-77.10
-77.05
-77.00
-76.95
1
2

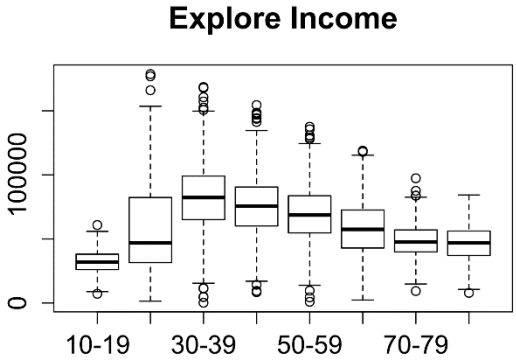

Sites for three kiosks
Latitude
Longitude
39.00
38.95
38.90
38.85
-77.10
-77.05
-77.00
-76.95
1
2
3

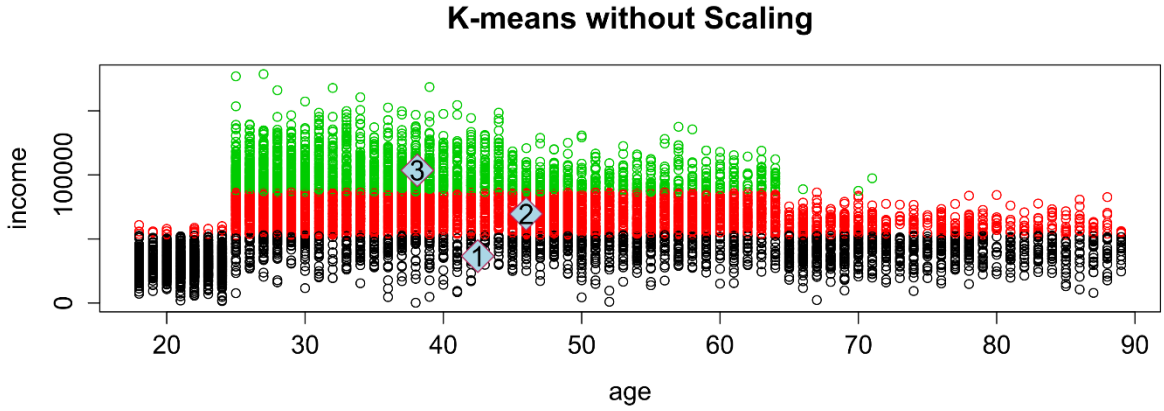

Hybrid: Two-cluster kiosks in three-cluster locations
Latitude
Longitude
39.00
38.95
38.90
38.85
-77.10
-77.05
-77.00
-76.95
1
2

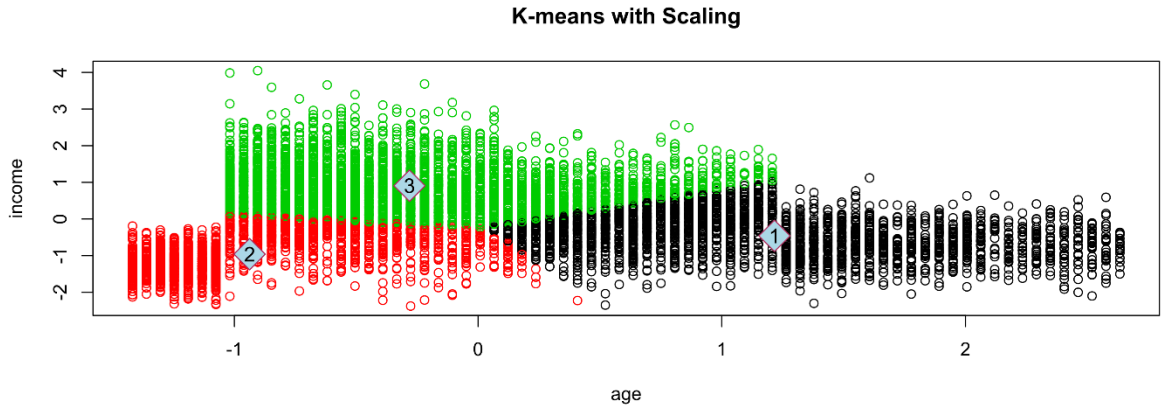

Two Kiosks
Three Kiosks
Two (Hybrid Solution)
Frequency
Dist (km)
0
20
40
60
80
0
2
4
6
8

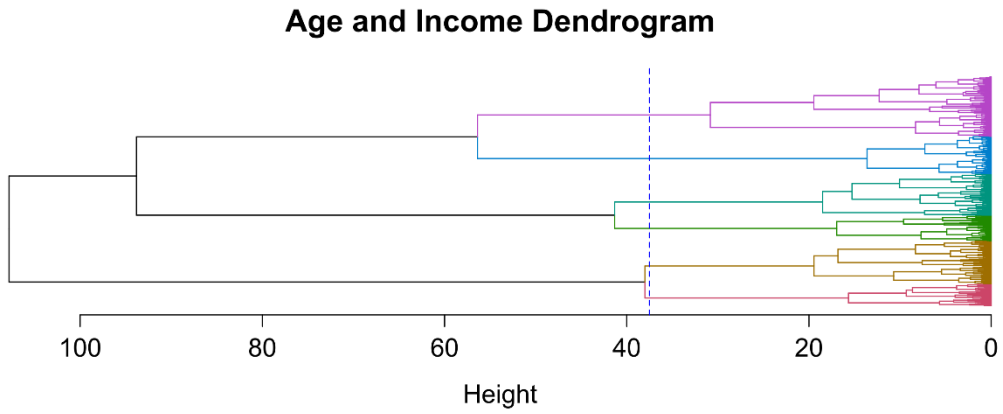

Distance Increase: Building Two Kiosks at Future Locations
Frequency
0
50
100
150
Average increase: 1.42 miles
152
14
22
35
13
8
-1
0
1
2
3
4
5
Distance (km)

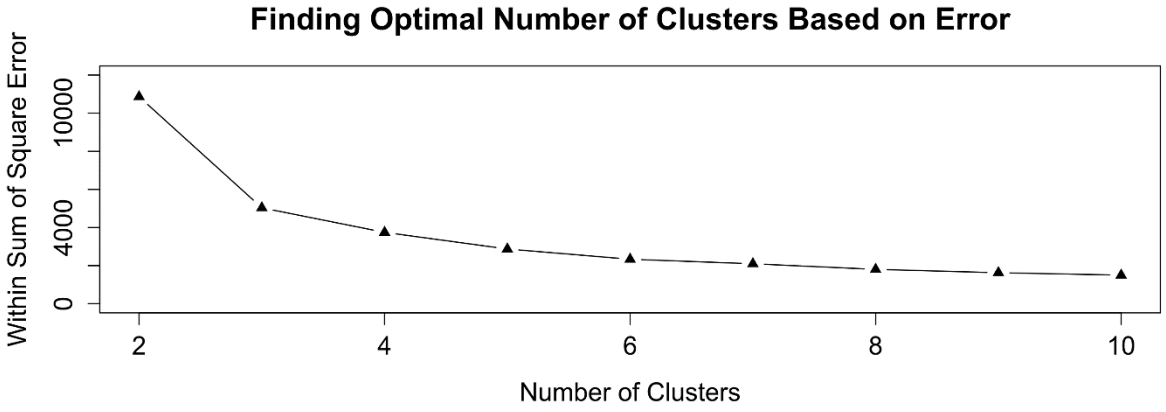

Explore Age
20
40
60
80
10-19
30-39
50-59
70-79

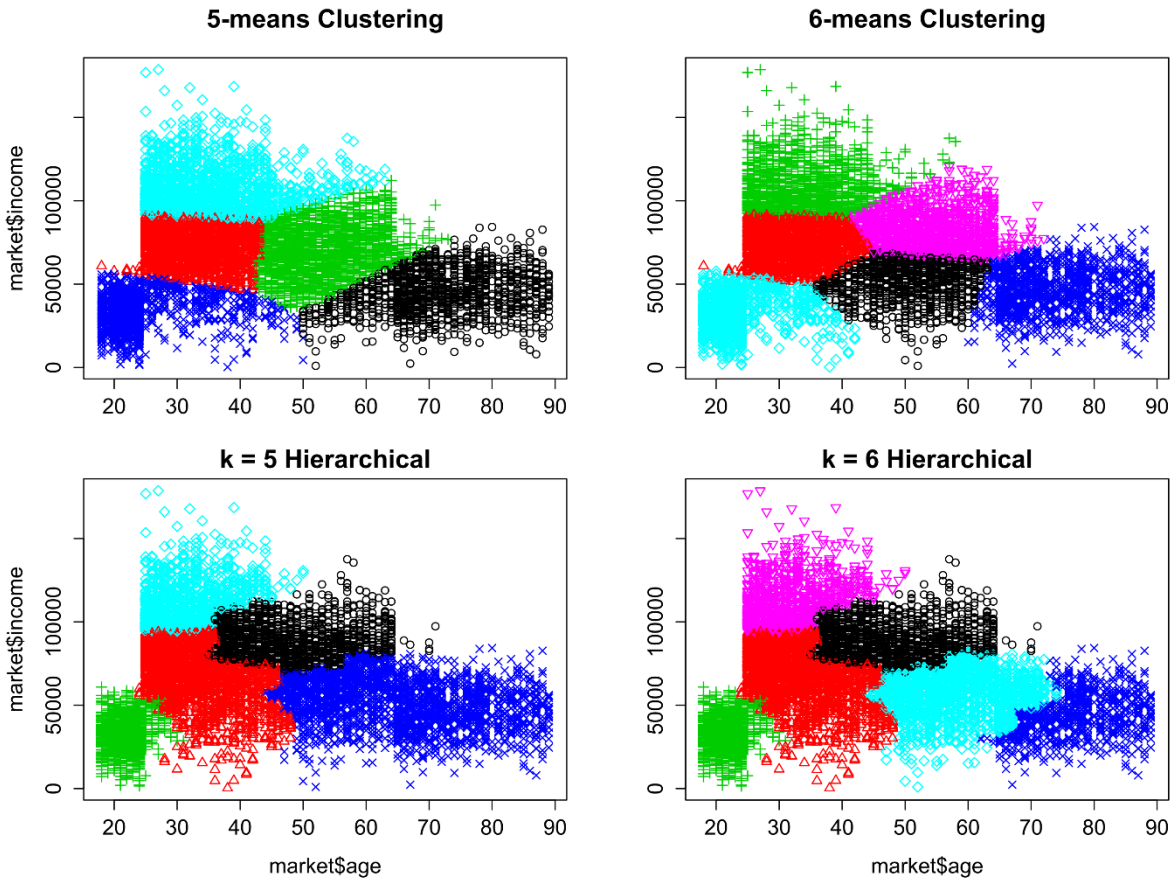

Explore Income
0
100000
10-19
30-39
50-59
70-79

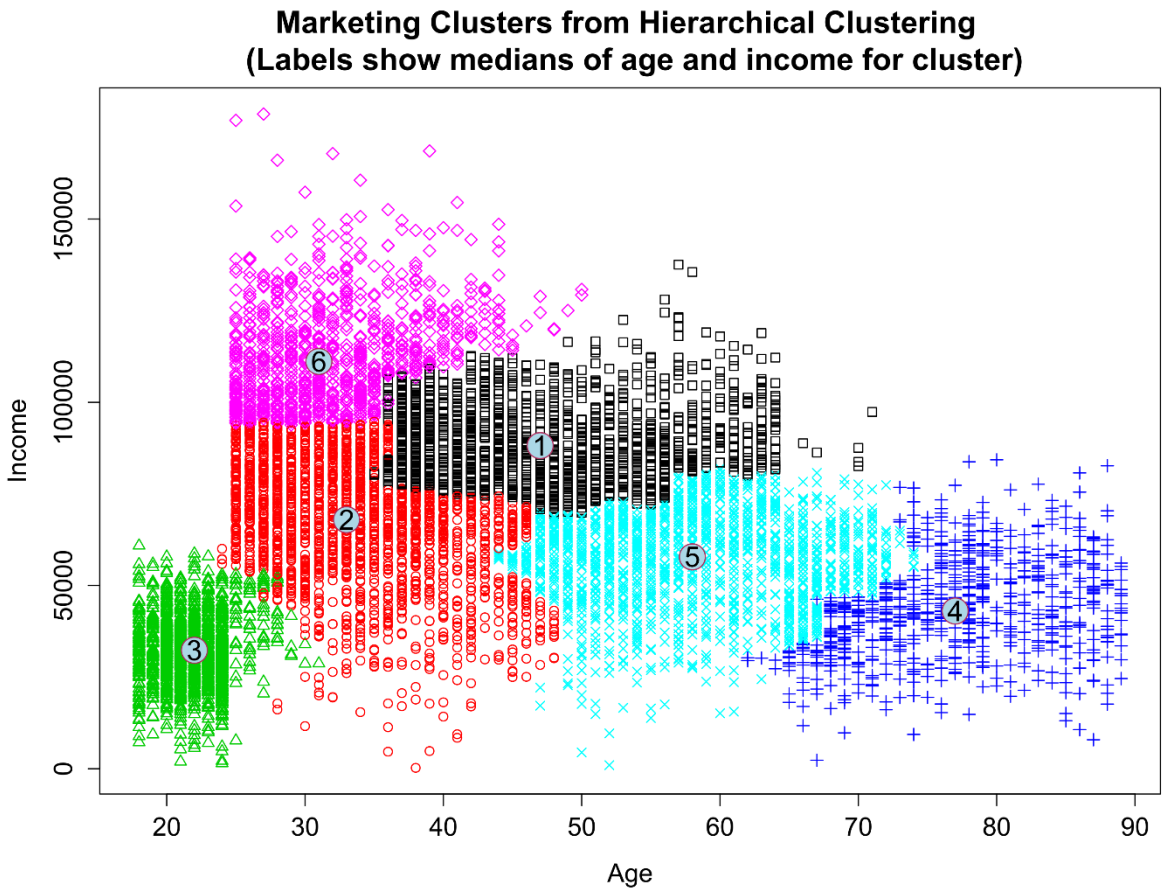

K-means without Scaling
income
0
100000
3
2
1
20
30
40
50
60
70
80
90
age


K-means with Scaling
income
age
-2
-1
0
1
2
3
4
-1
0
1
2
1
2
3

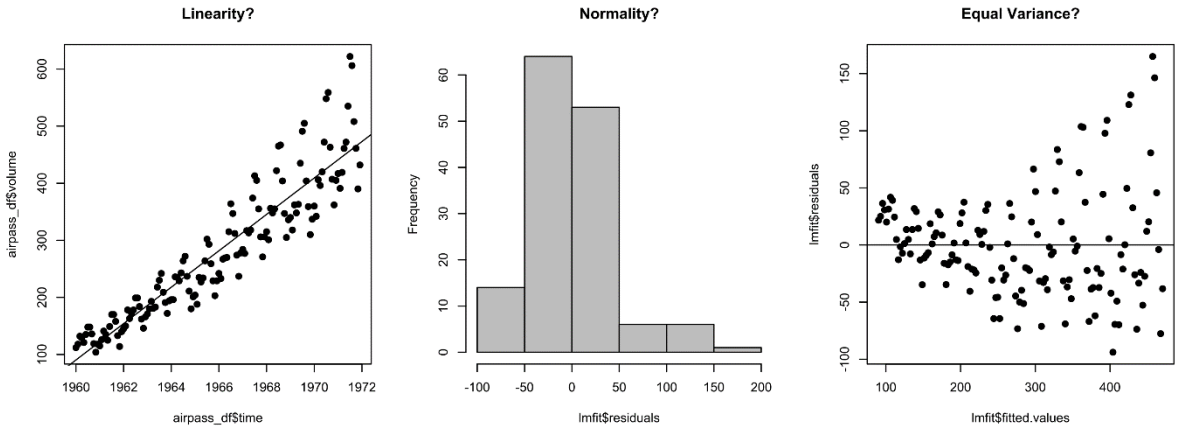

Age and Income Dendrogram
100
80
60
40
20
0
Height

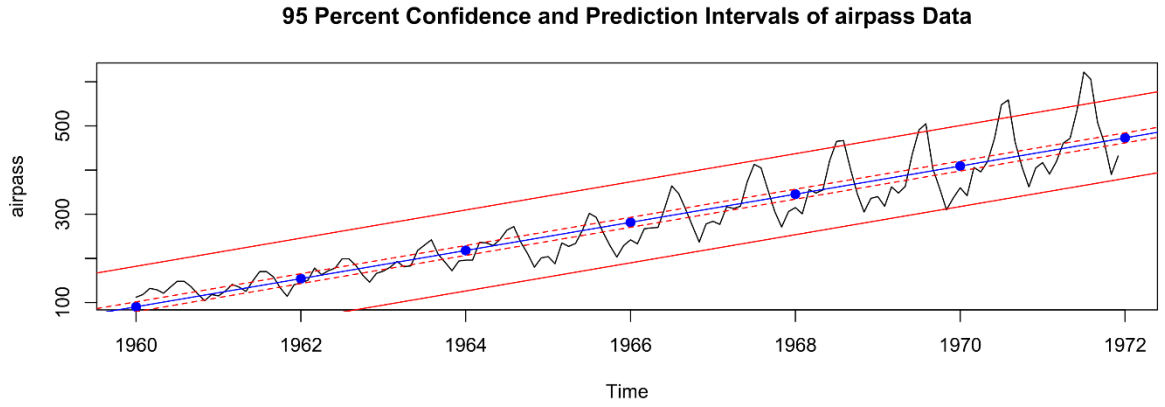

Finding Optimal Number of Clusters Based on Error
Within Sum of Square Error
0
4000
10000
2
4
6
8
10
Number of Clusters

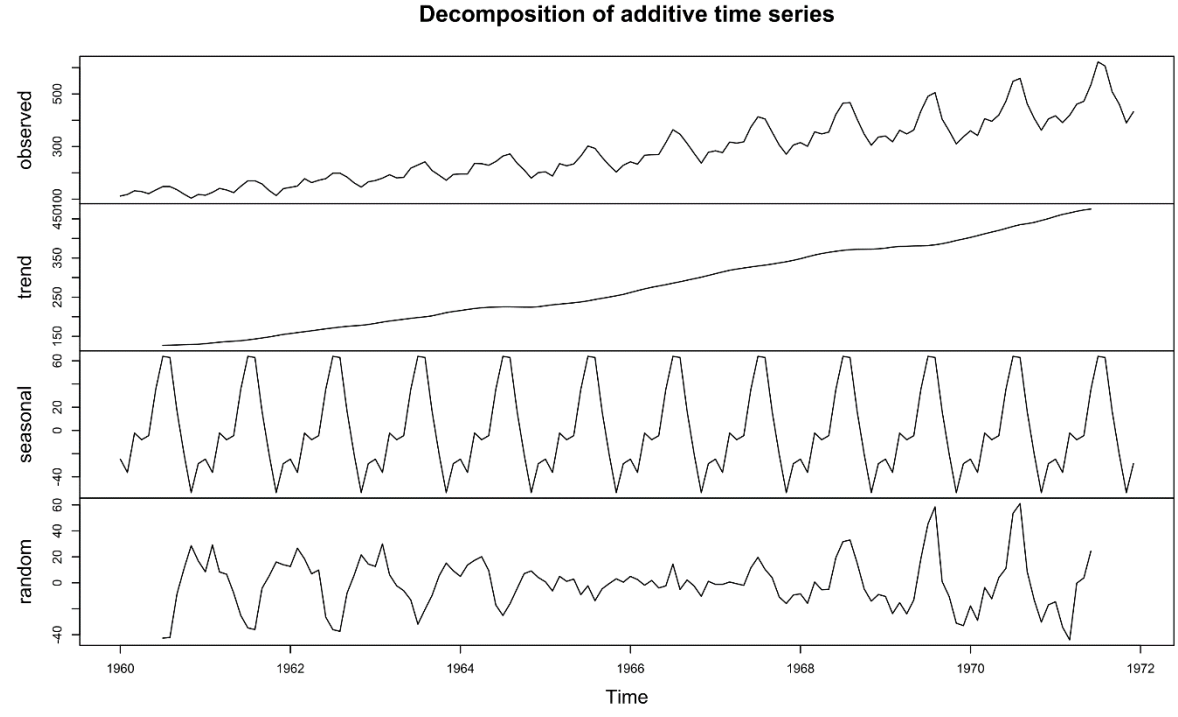

5-means Clustering
6-means Clustering
k = 5 Hierarchical
k = 6 Hierarchical
market$income
market$age
0
50000
100000
20
30
40
50
60
70
80
90

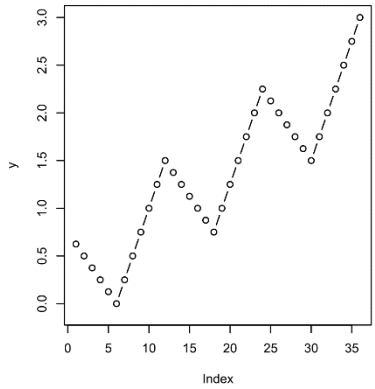

Marketing Clusters from Hierarchical Clustering
(Labels show medians of age and income for cluster)
Income
0
50000
100000
150000
20
30
40
50
60
70
80
90
Age
1
2
3
4
5
6

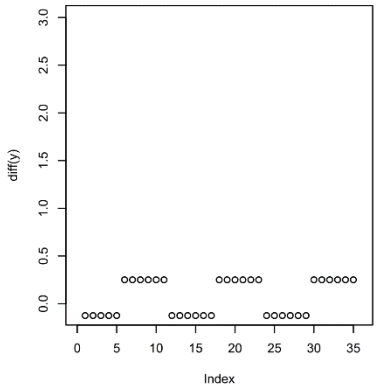
Chapter 6: Time Series Analysis

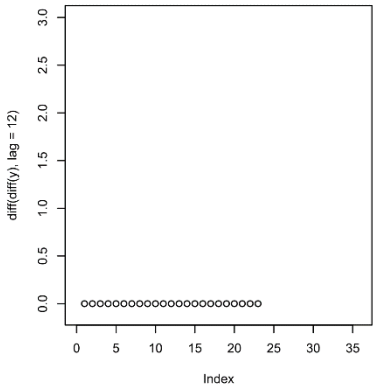




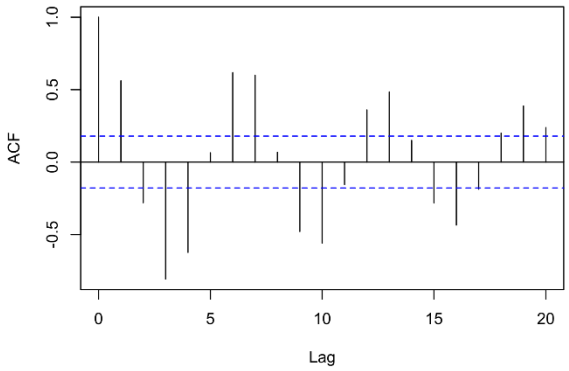




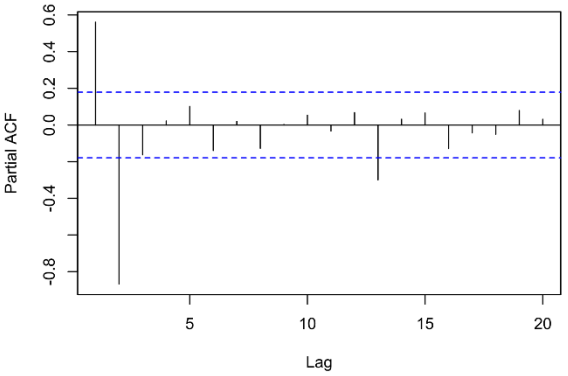




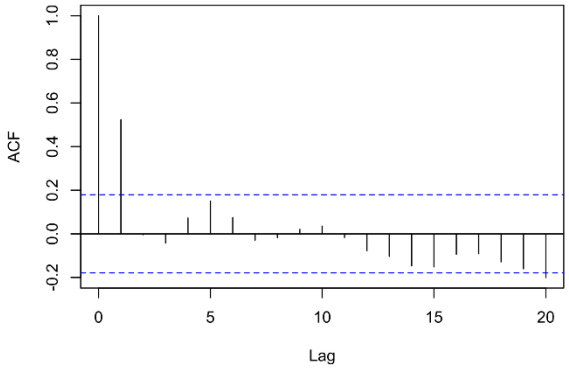
AR(2) Model



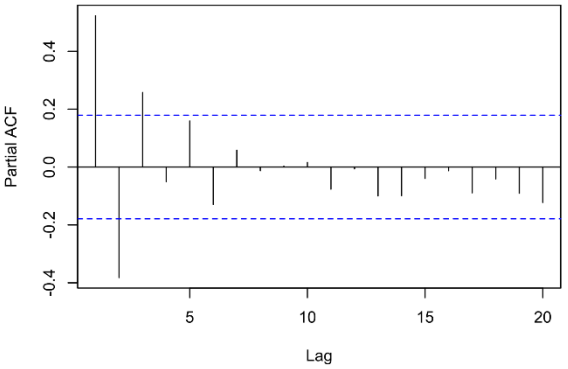
AR(2) Model

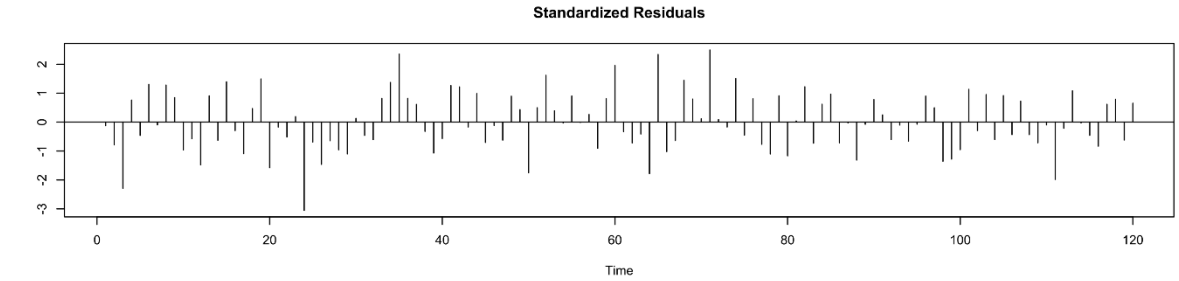


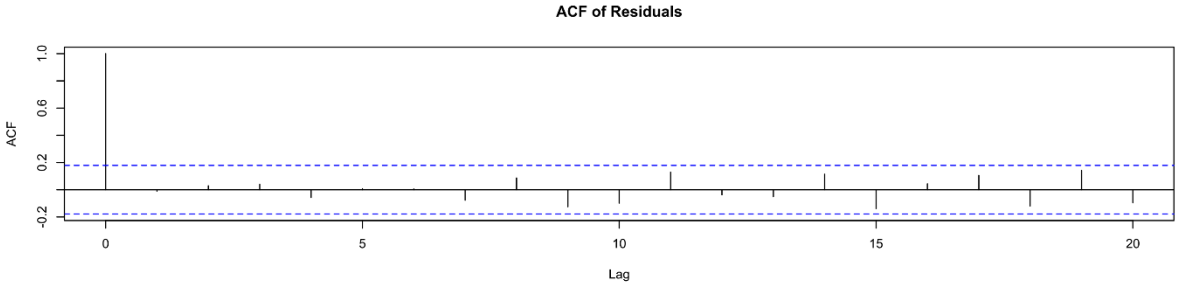
MA(1) Model

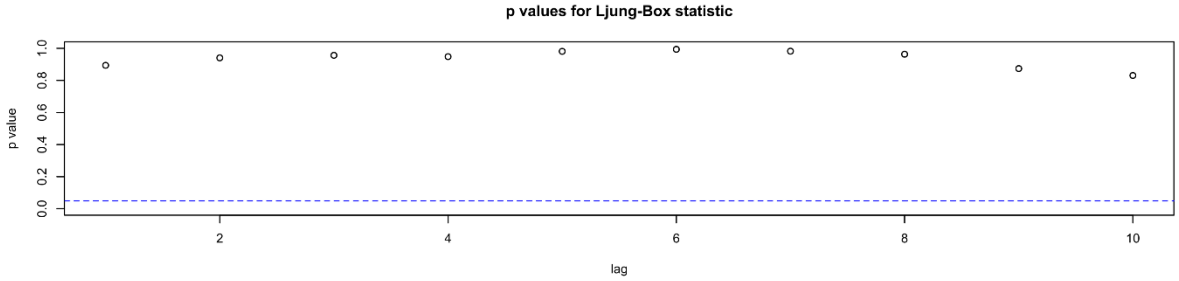


MA(1) Model

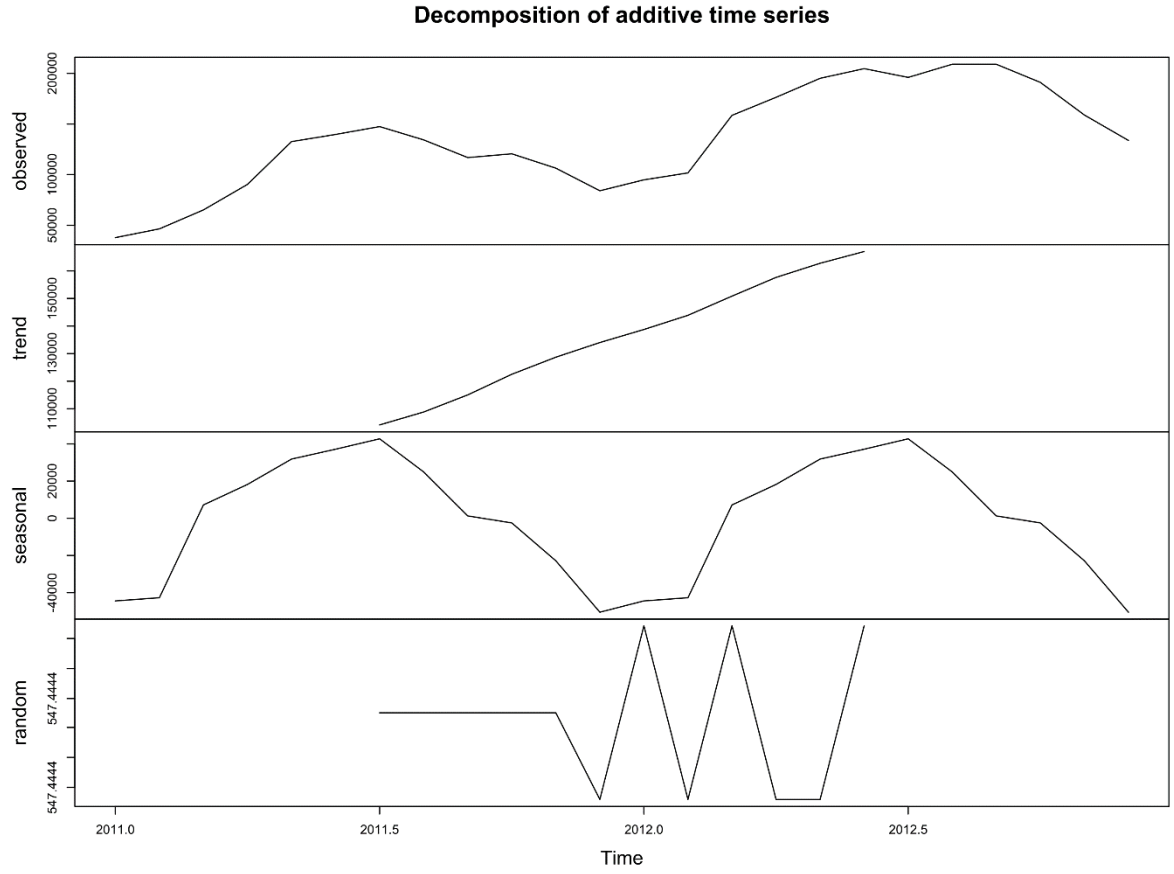


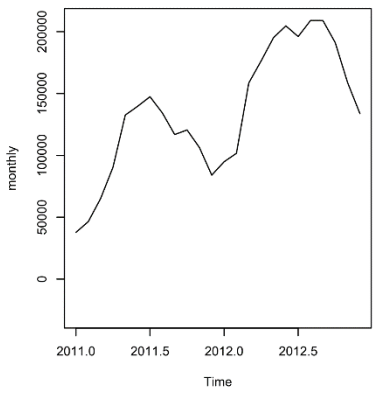


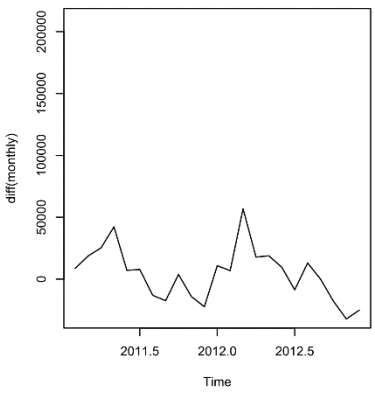


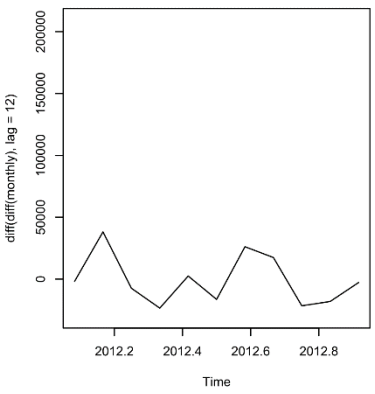


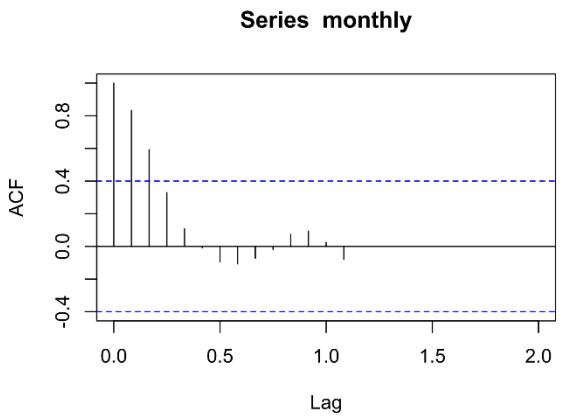

```
> monthly
      Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2011  37727  46396  65109  90332 132580 139674 147426 134280 116825 120535 106361  84025
2012  94832 101668 158535 176349 195114 204683 196014 209024 208995 191108 158855 133735
```

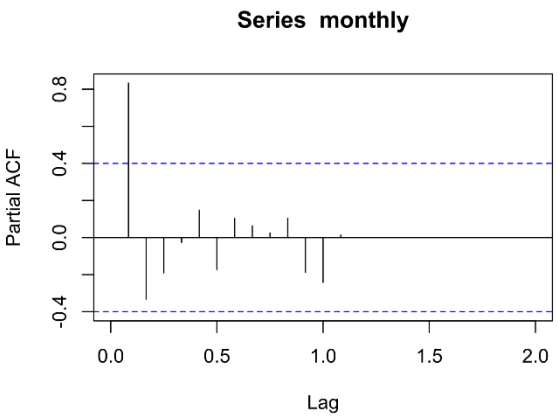

Decomposition of additive time series
observed
trend
seasonal
random
Time

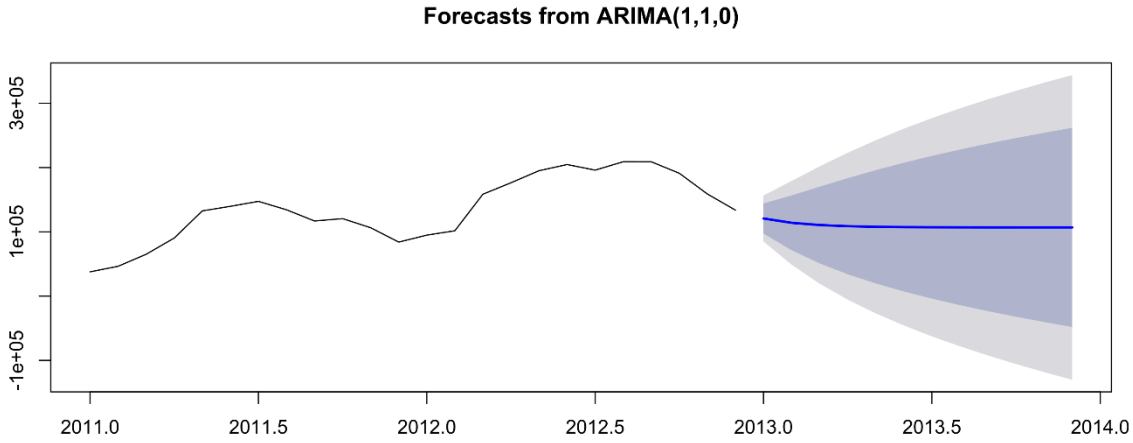

monthly
Time

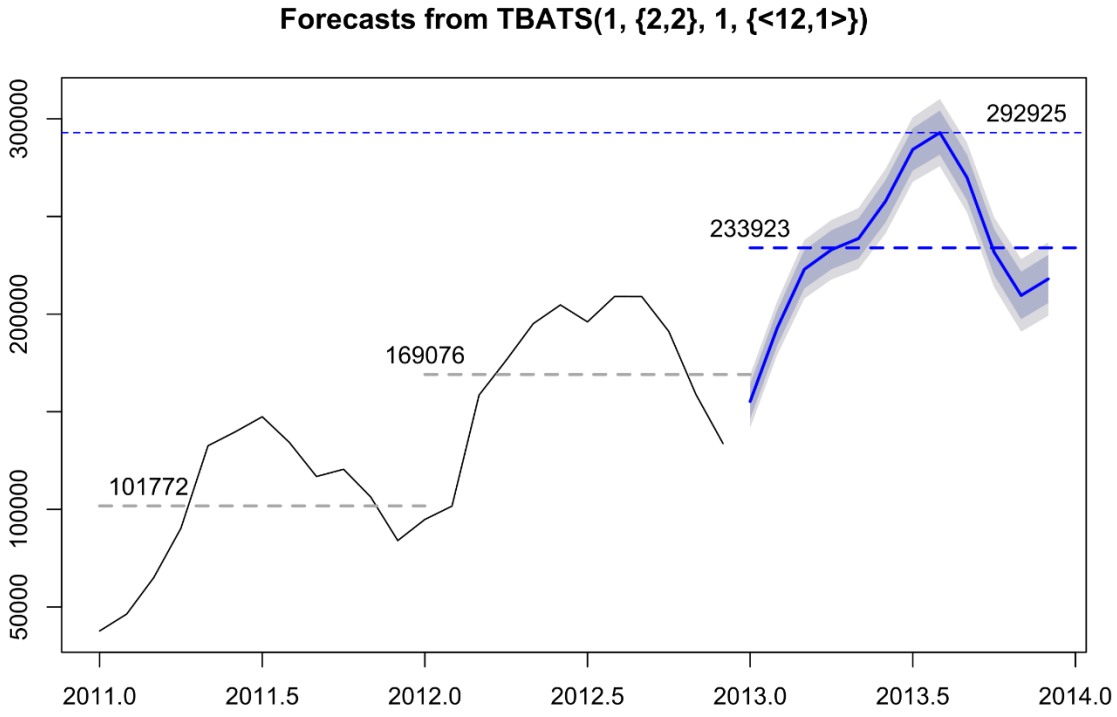

diff(monthly)
Time


diff(diff(monthly), lag = 12)
Time

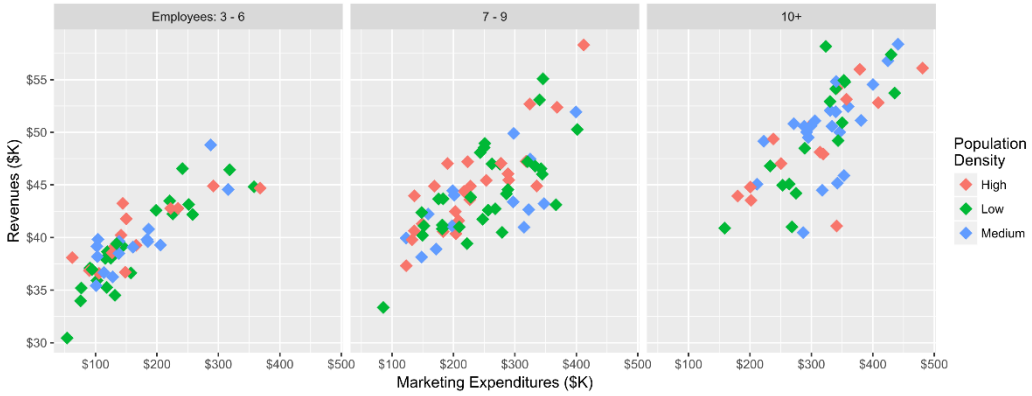

Series monthly
ACF
0.8
0.4
0.0
-0.4
0.0
0.5
1.0
1.5
2.0
Lag

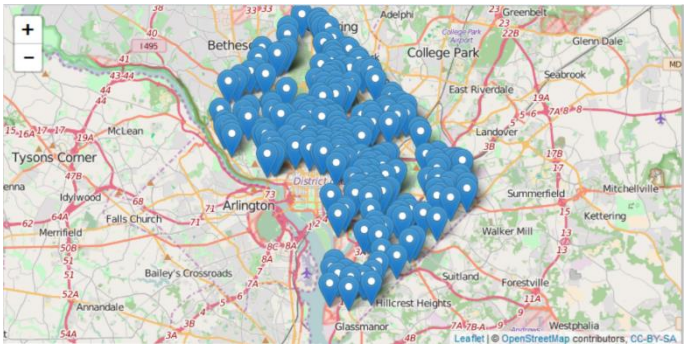

Series monthly
Partial ACF
0.8
0.4
0.0
-0.4
0.0
0.5
1.0
1.5
2.0
Lag

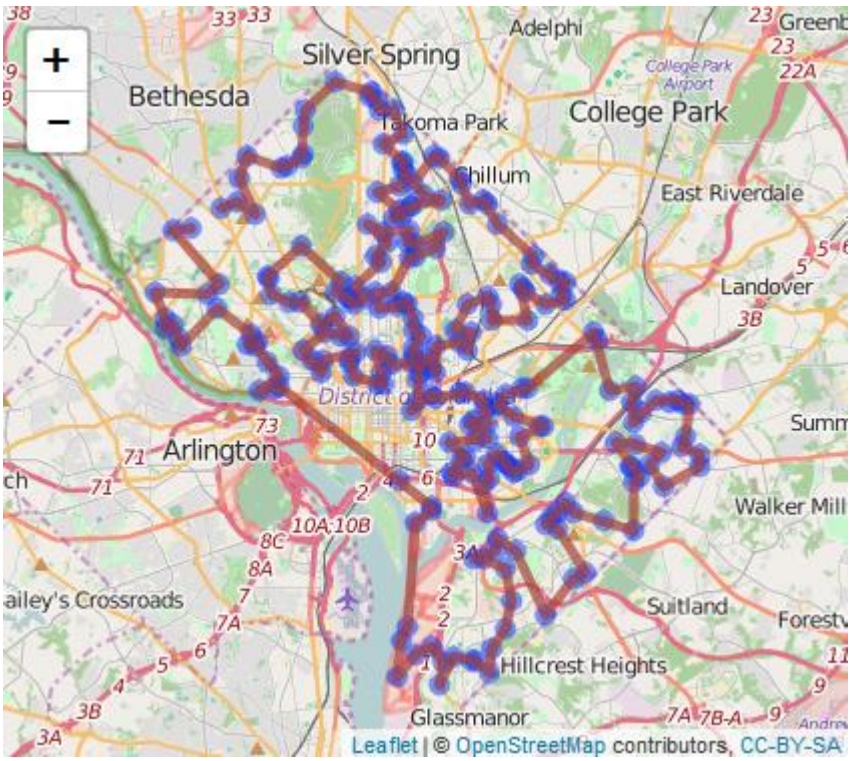

Forecasts from ARIMA(1,1,0)
3e+05
1e+05
-1e+05
2011.0
2011.5
2012.0
2012.5
2013.0
2013.5
2014.0


Forecasts from TBATS(1, {2,2}, 1, {<12,1>})
300000
200000
100000
50000
292925
233923
169076
101772
2011.0
2011.5
2012.0
2012.5
2013.0
2013.5
2014.0

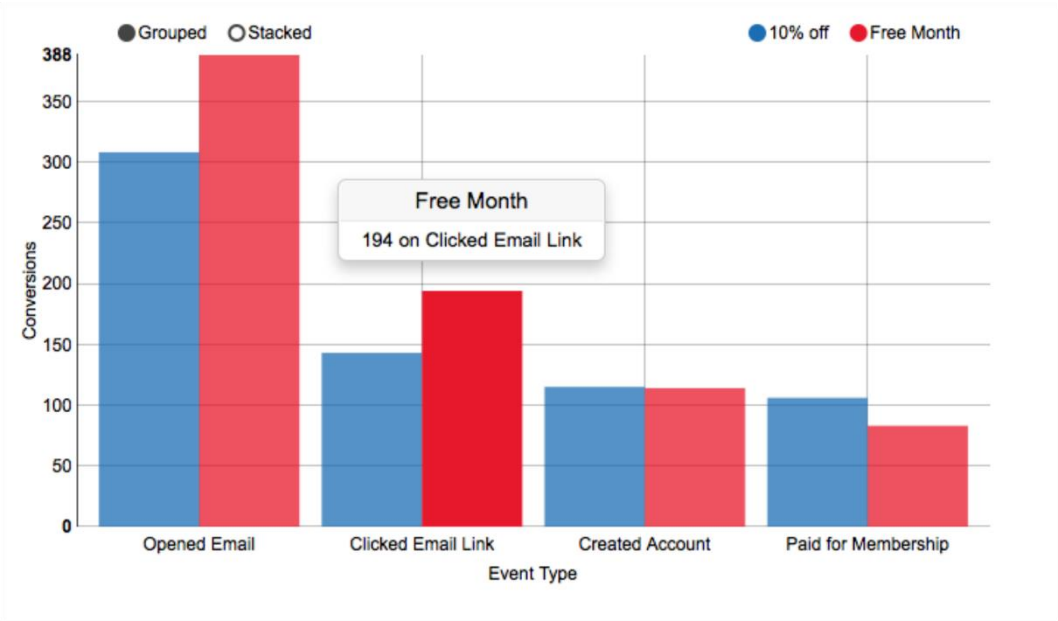
Chapter 7: Visualizing the Data's Story

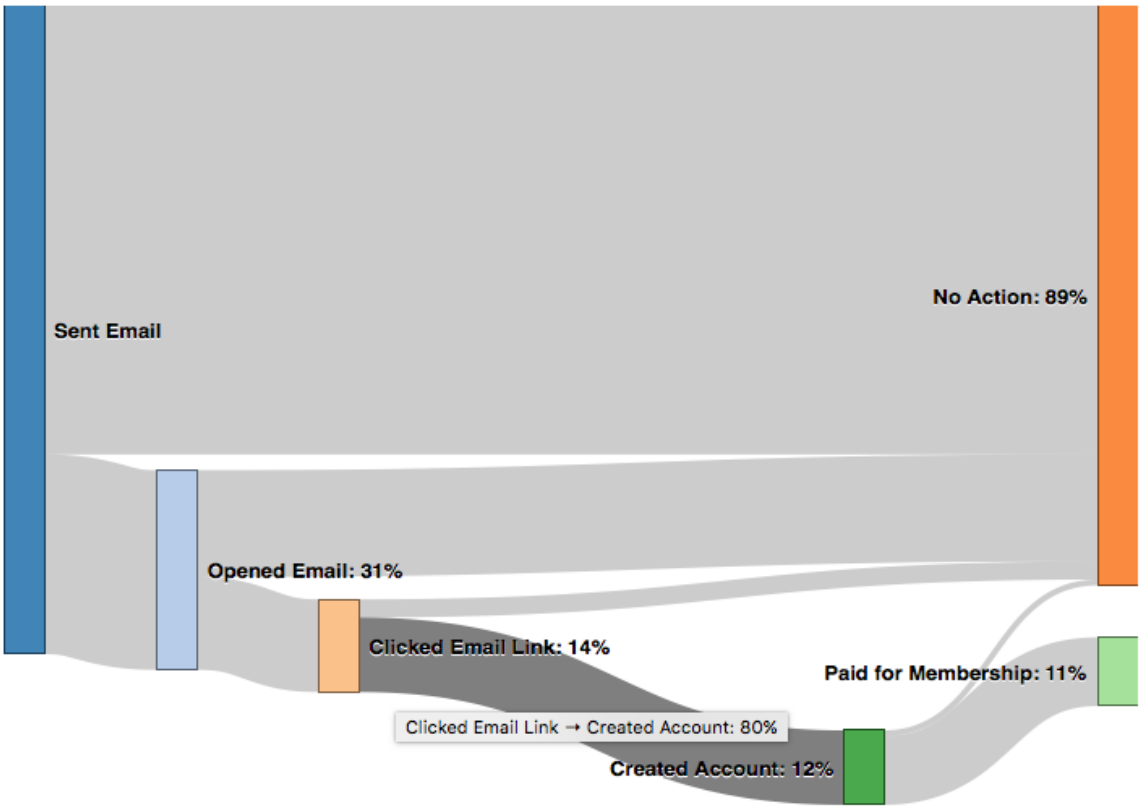




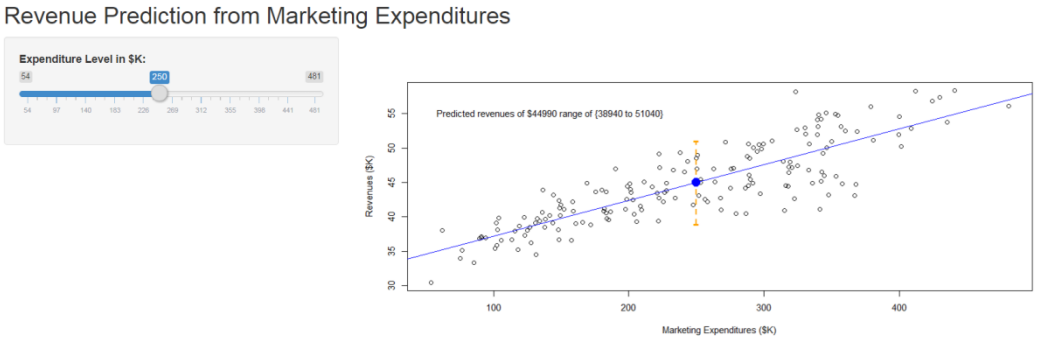


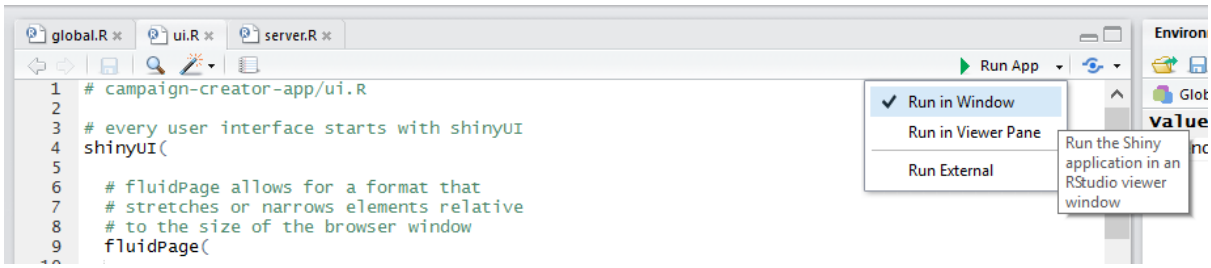
	Consumer	Promotion	Opened Email	Clicked Email Link	Created Account	Paid for Membership
1	1	Free Month	N	N	N	N
2	2	10% off	N	N	N	N
3	3	Free Month	N	N	N	N
4	4	10% off	N	N	N	N
5	5	Free Month	N	N	N	N
6	6	10% off	N	N	N	N

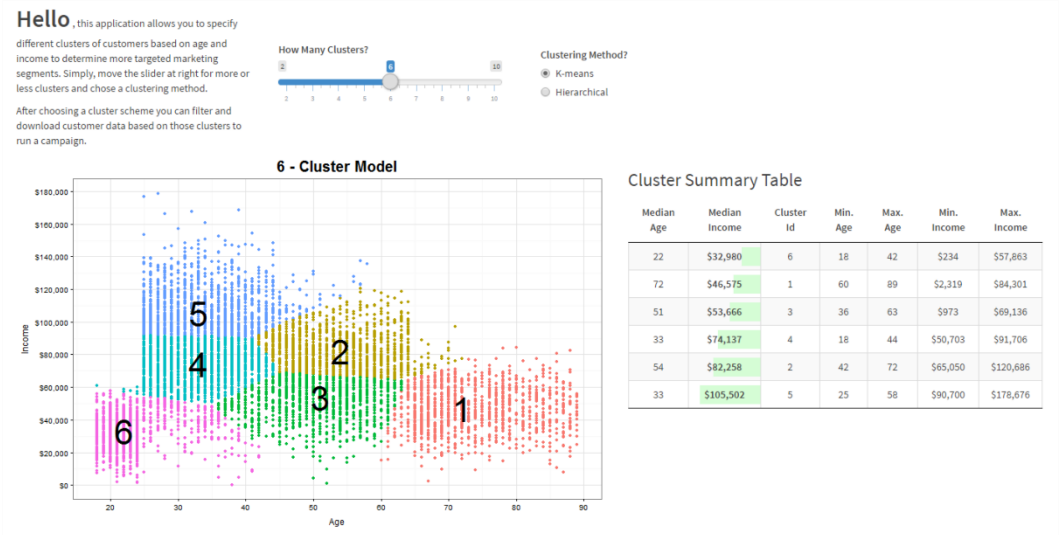

Grouped
Stacked
10% off
Free Month
388
350
300
250
200
150
100
50
0
Conversions
Free Month
194 on Clicked Email Link
Opened Email
Clicked Email Link
Created Account
Paid for Membership
Event Type


Sent Email
No Action: 89%
Opened Email: 31%
Clicked Email Link: 14%
Paid for Membership: 11%
Clicked Email Link → Created Account: 80%
Created Account: 12%

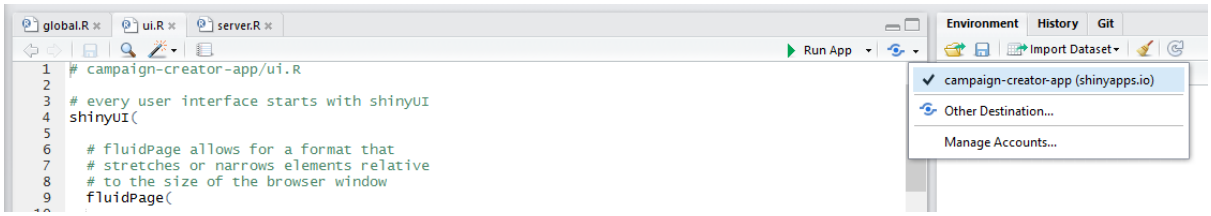
Chapter 8: Web Dashboards with Shiny







Median Age	Median Income	Cluster Id	Min. Age	Max. Age	Min. Income	Max. Income
22	$32,980	6	18	42	$234	$57,863
72	$46,575	1	60	89	$2,319	$84,301
51	$53,666	3	36	63	$973	$69,136
33	$74,137	4	18	44	$50,703	$91,706
54	$82,258	2	42	72	$65,050	$120,686
33	$105,502	5	25	58	$90,700	$178,676


global.R
ui.R
server.R
Run App
Environment
History
Git
Import Dataset
campaign-creator-app (shinyapps.io)
Other Destination...
Manage Accounts...
1 # campaign-creator-app/ui.R
2
3 # every user interface starts with shinyUI
4 shinyUI(
5
6 # fluidPage allows for a format that
7 # stretches or narrows elements relative
8 # to the size of the browser window
9 fluidPage(