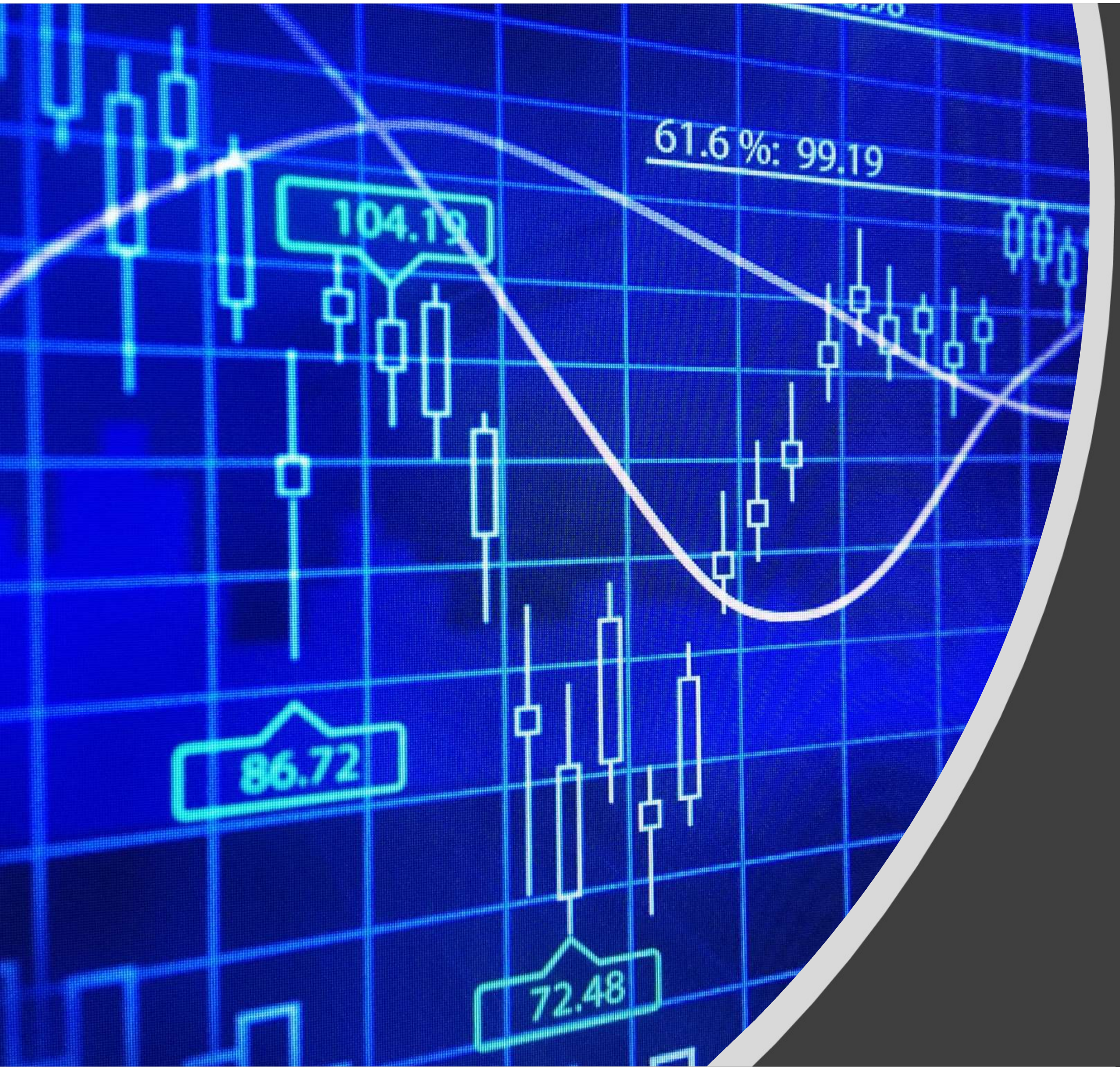
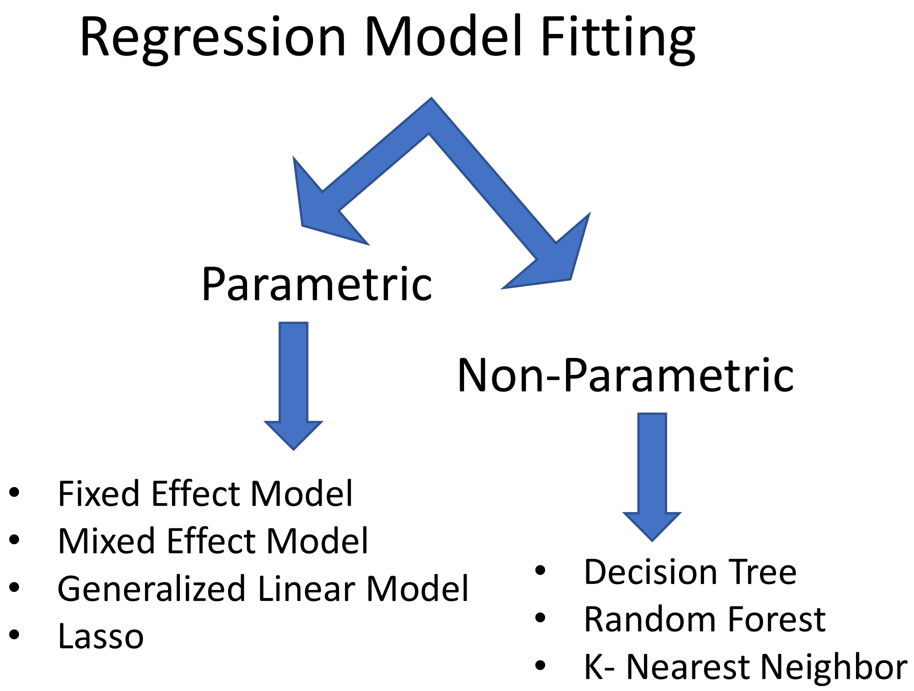


Sales Data Analysis

Rochita Das

Contents

- Snapshot of Data
- Data Description & Preparation
- Exploratory Data Analysis
- Regression Model Fitting
- Time Series Analysis
- Market Basket Analysis
- Future Scope



Regression Model Fitting

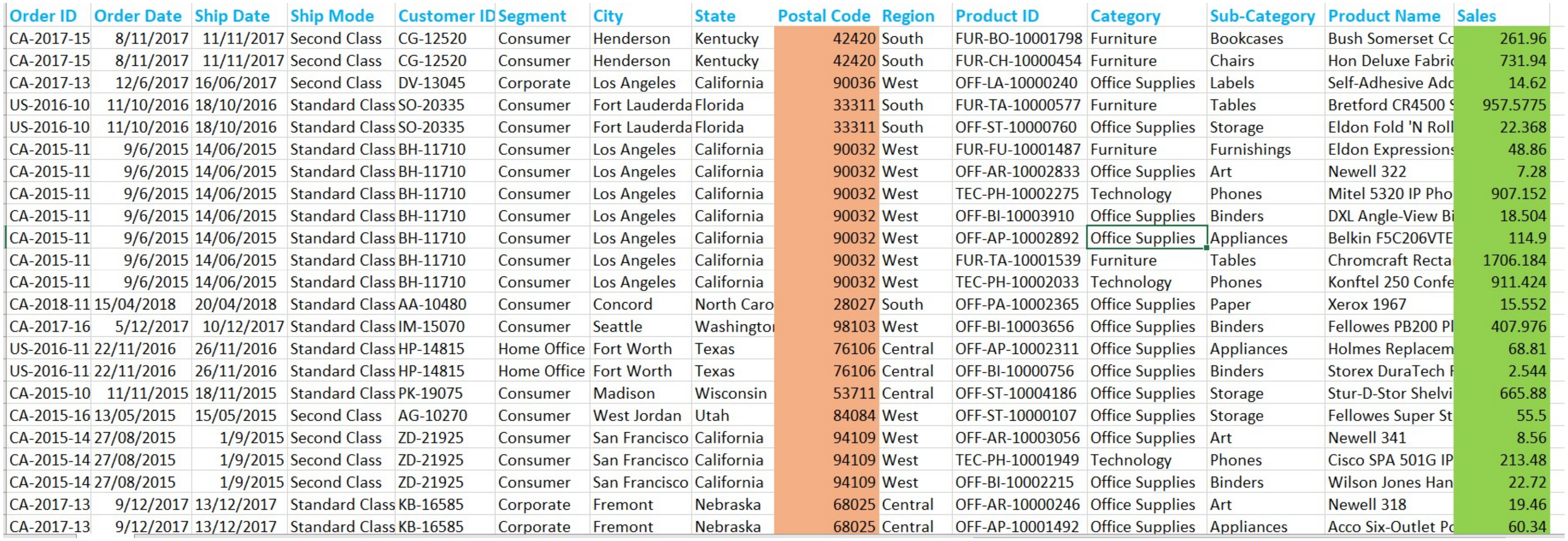
Parametric

- Fixed Effect Model
- Mixed Effect Model
- Generalized Linear Model
- Lasso

Non-Parametric

- Decision Tree
- Random Forest
- K- Nearest Neighbor

Snapshot of Data

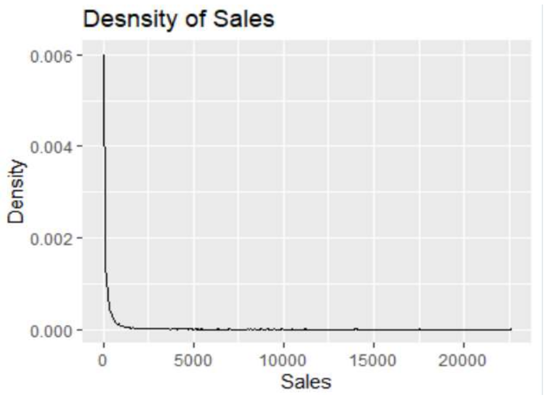


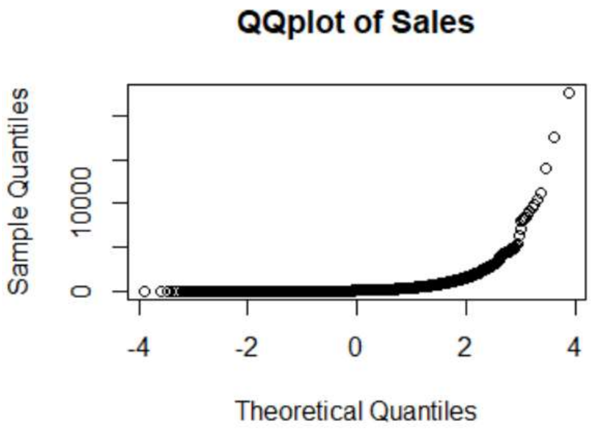
Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Segment	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales
CA-2017-15	8/11/2017	11/11/2017	Second Class	CG-12520	Consumer	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset C	261.96
CA-2017-15	8/11/2017	11/11/2017	Second Class	CG-12520	Consumer	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabri	731.94
CA-2017-13	12/6/2017	16/06/2017	Second Class	DV-13045	Corporate	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Ad	14.62
US-2016-10	11/10/2016	18/10/2016	Standard Class	SO-20335	Consumer	Fort Lauderda	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretford CR4500	957.5775
US-2016-10	11/10/2016	18/10/2016	Standard Class	SO-20335	Consumer	Fort Lauderda	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll	22.368
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon Expressions	48.86
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell 322	7.28
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5320 IP Pho	907.152
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Angle-View Bi	18.504
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C206VTE	114.9
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	FUR-TA-10001539	Furniture	Tables	Chromcraft Recta	1706.184
CA-2015-11	9/6/2015	14/06/2015	Standard Class	BH-11710	Consumer	Los Angeles	California	90032	West	TEC-PH-10002033	Technology	Phones	Konftel 250 Confe	911.424
CA-2018-11	15/04/2018	20/04/2018	Standard Class	AA-10480	Consumer	Concord	North Caro	28027	South	OFF-PA-10002365	Office Supplies	Paper	Xerox 1967	15.552
CA-2017-16	5/12/2017	10/12/2017	Standard Class	IM-15070	Consumer	Seattle	Washingto	98103	West	OFF-BI-10003656	Office Supplies	Binders	Fellowes PB200 Pl	407.976
US-2016-11	22/11/2016	26/11/2016	Standard Class	HP-14815	Home Office	Fort Worth	Texas	76106	Central	OFF-AP-10002311	Office Supplies	Appliances	Holmes Replacem	68.81
US-2016-11	22/11/2016	26/11/2016	Standard Class	HP-14815	Home Office	Fort Worth	Texas	76106	Central	OFF-BI-10000756	Office Supplies	Binders	Storex DuraTech	2.544
CA-2015-10	11/11/2015	18/11/2015	Standard Class	PK-19075	Consumer	Madison	Wisconsin	53711	Central	OFF-ST-10004186	Office Supplies	Storage	Stur-D-Stor Shelvi	665.88
CA-2015-16	13/05/2015	15/05/2015	Second Class	AG-10270	Consumer	West Jordan	Utah	84084	West	OFF-ST-10000107	Office Supplies	Storage	Fellowes Super St	55.5
CA-2015-14	27/08/2015	1/9/2015	Second Class	ZD-21925	Consumer	San Francisco	California	94109	West	OFF-AR-10003056	Office Supplies	Art	Newell 341	8.56
CA-2015-14	27/08/2015	1/9/2015	Second Class	ZD-21925	Consumer	San Francisco	California	94109	West	TEC-PH-10001949	Technology	Phones	Cisco SPA 501G IP	213.48
CA-2015-14	27/08/2015	1/9/2015	Second Class	ZD-21925	Consumer	San Francisco	California	94109	West	OFF-BI-10002215	Office Supplies	Binders	Wilson Jones Han	22.72
CA-2017-13	9/12/2017	13/12/2017	Standard Class	KB-16585	Corporate	Fremont	Nebraska	68025	Central	OFF-AR-10000246	Office Supplies	Art	Newell 318	19.46
CA-2017-13	9/12/2017	13/12/2017	Standard Class	KB-16585	Corporate	Fremont	Nebraska	68025	Central	OFF-AP-10001492	Office Supplies	Appliances	Acco Six-Outlet P	60.34

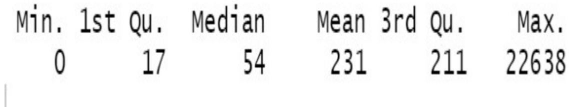
Data Description & Preparation

- N = 9800, No of Variables = 16
- Response variable : 'Sales'
- 'Order Date', 'Ship Date' – vary with time
- Year, Month and Weekdays are extracted from 'Order Date' - Categories
- Rest all variables are Categorical : Fixed Effect, Random Effect
- 'Zip code' has missing value– Imputed with State and City code
- Train – 75% data, Test – 25% data

Exploratory Data Analysis

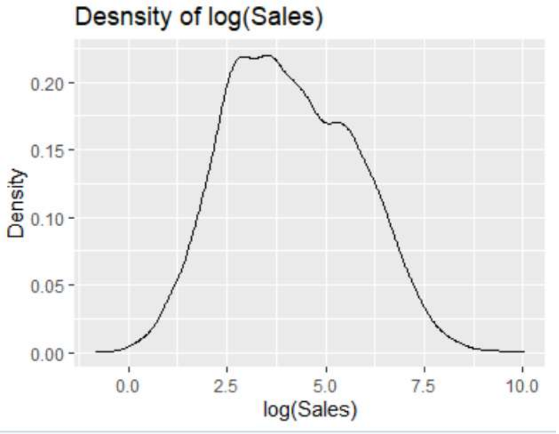


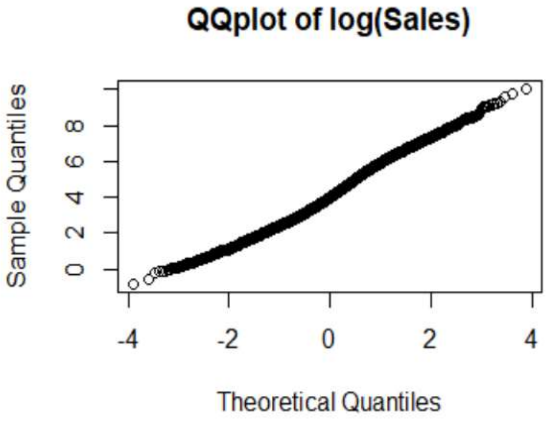


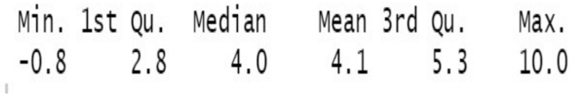


```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0      17      54     231     211   22638
```

Transformation : Log (Sales)

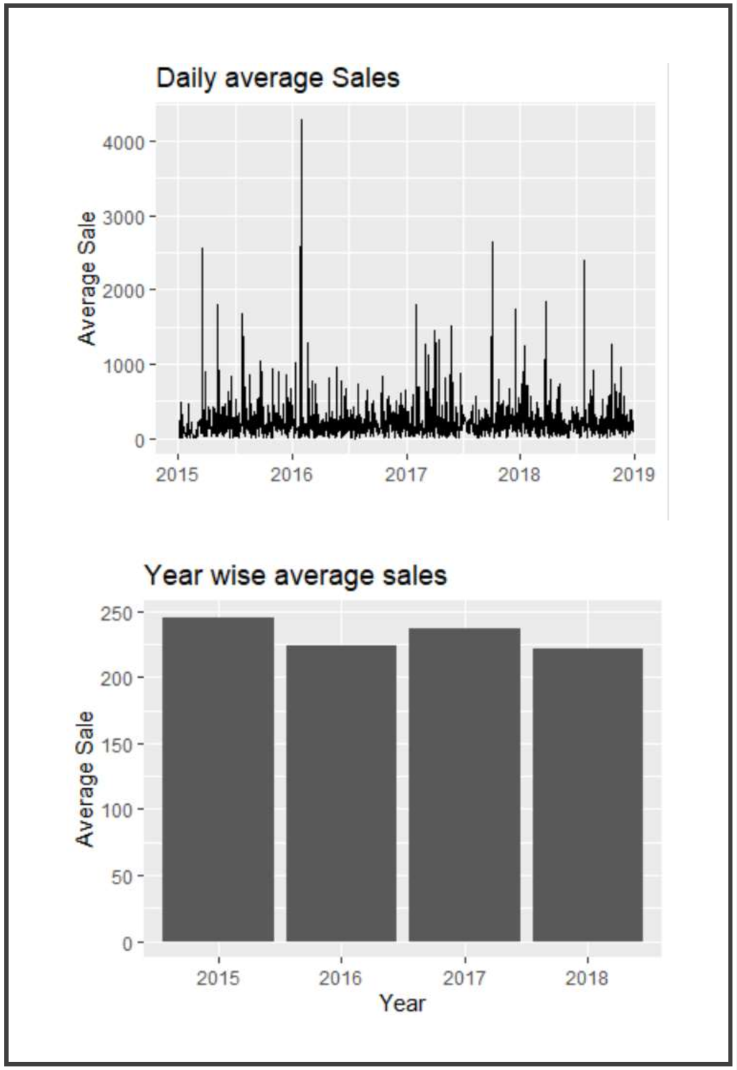


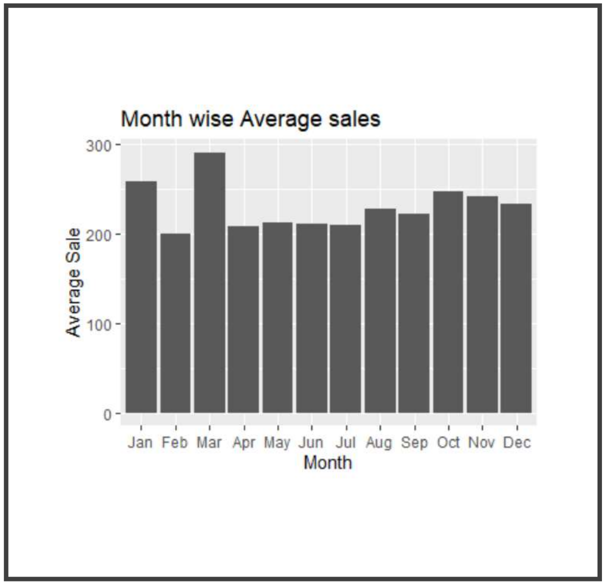


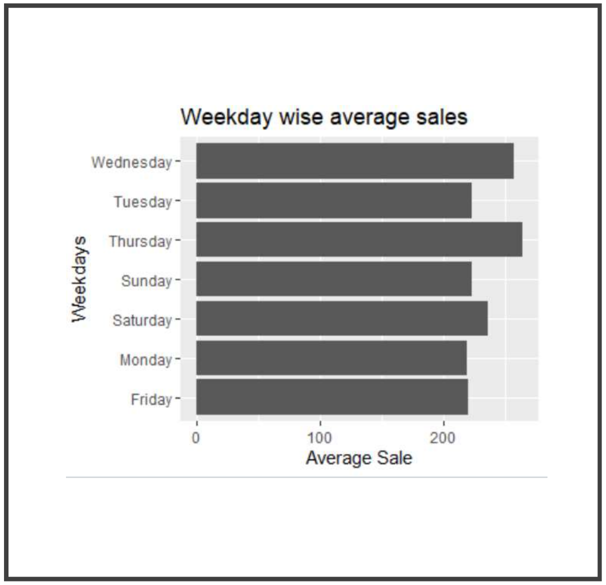


```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.8     2.8     4.0     4.1     5.3    10.0
```

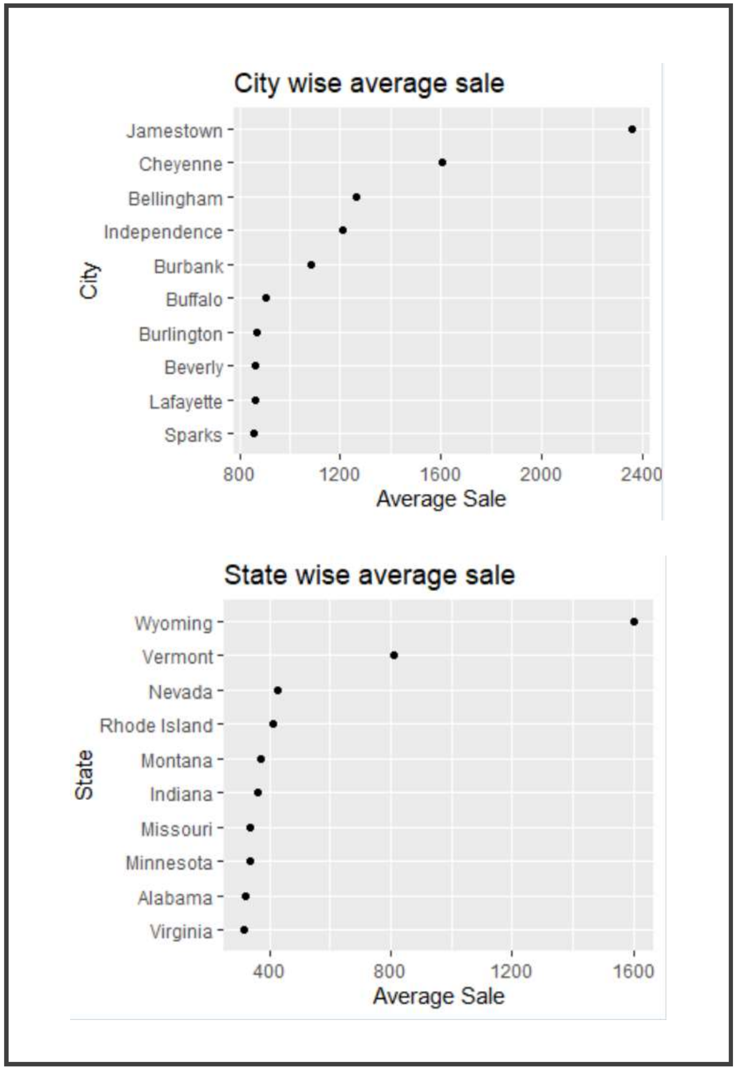
Exploratory Data Analysis





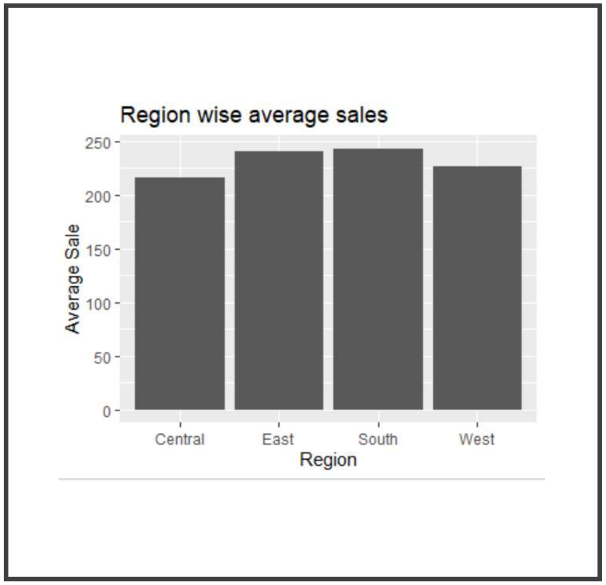


Exploratory Data Analysis

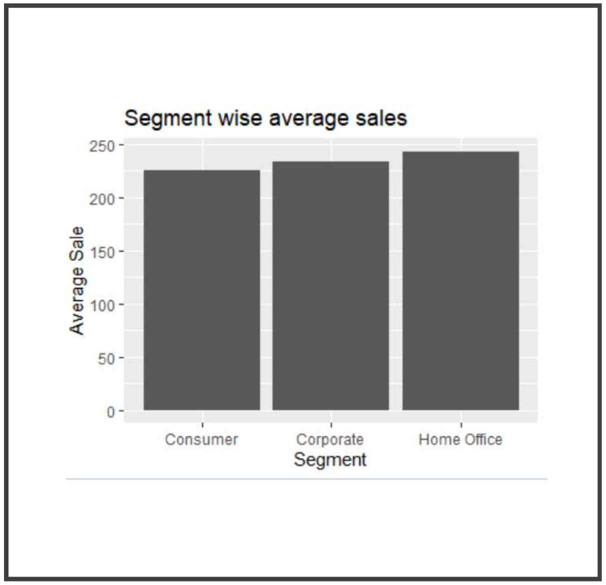


City wise average sale

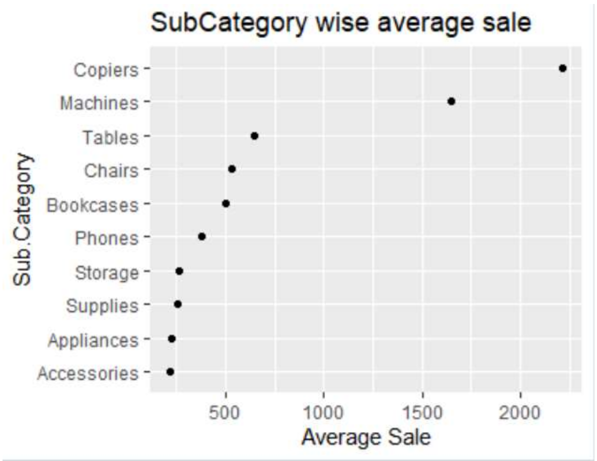
State wise average sale

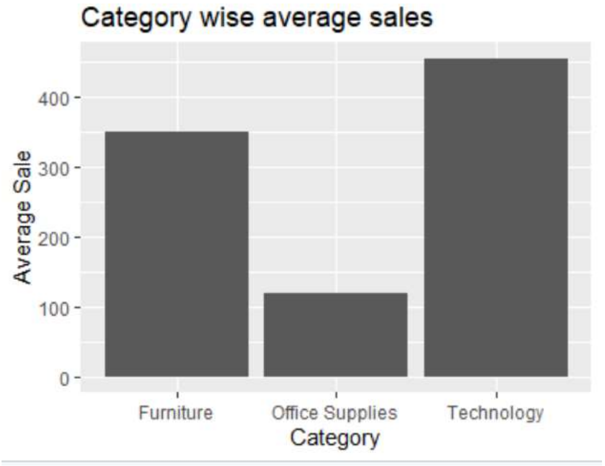


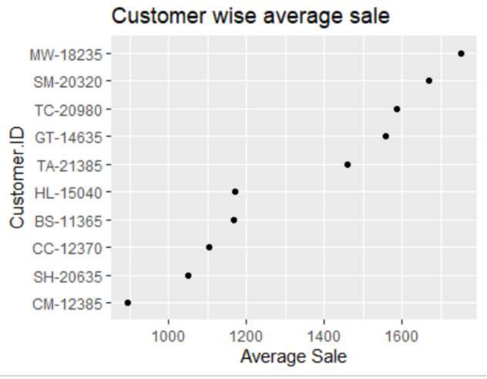
Region wise average sales

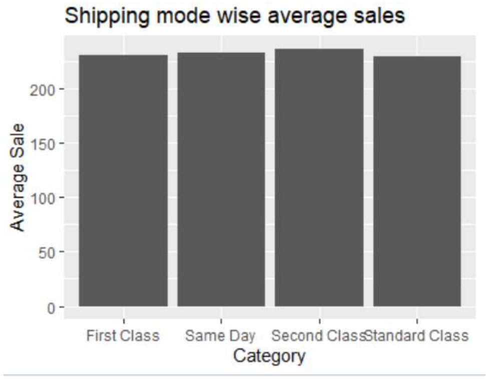


Segment wise average sales







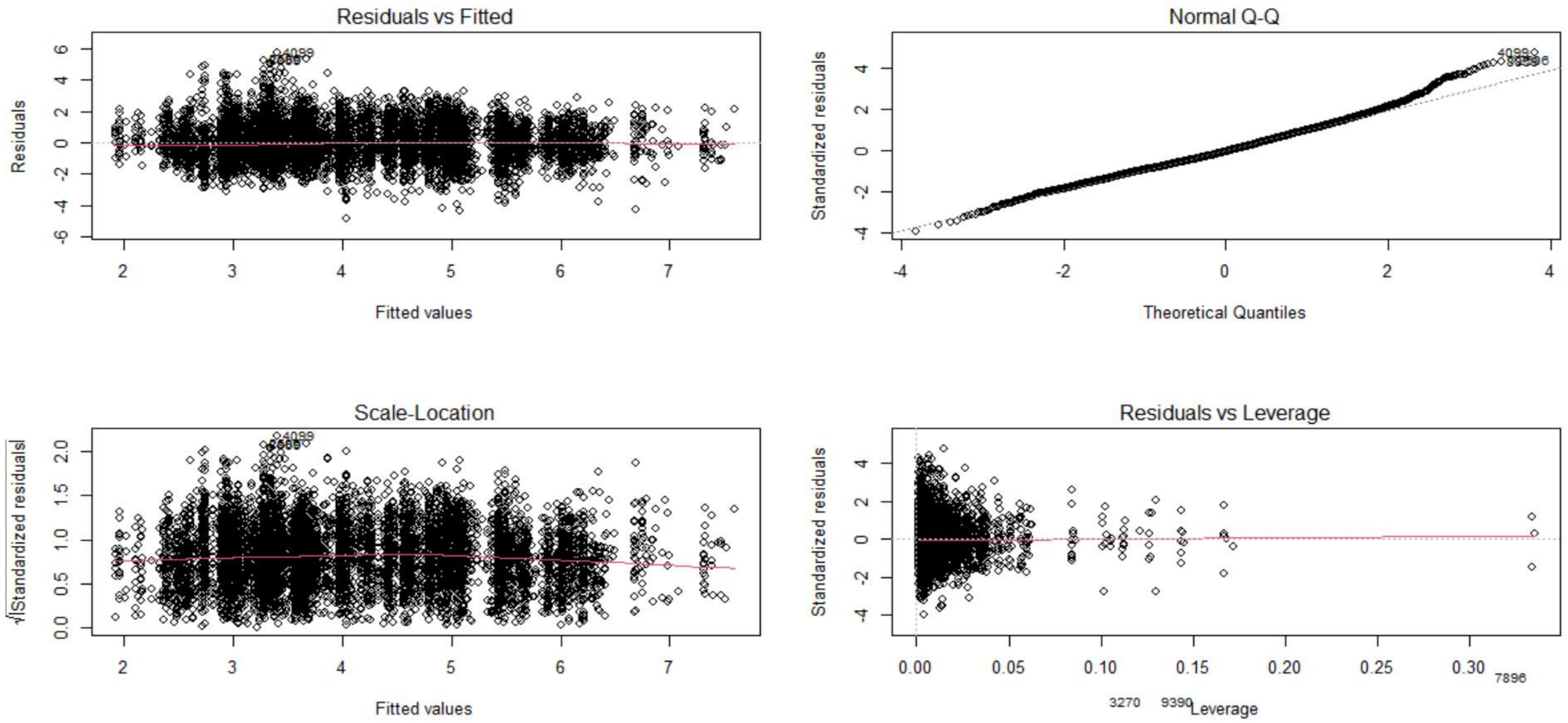


Exploratory Data Analysis

Regression Model Fitting: Fixed Effect

Model : E[Log(sales)]= States + Sub.Category

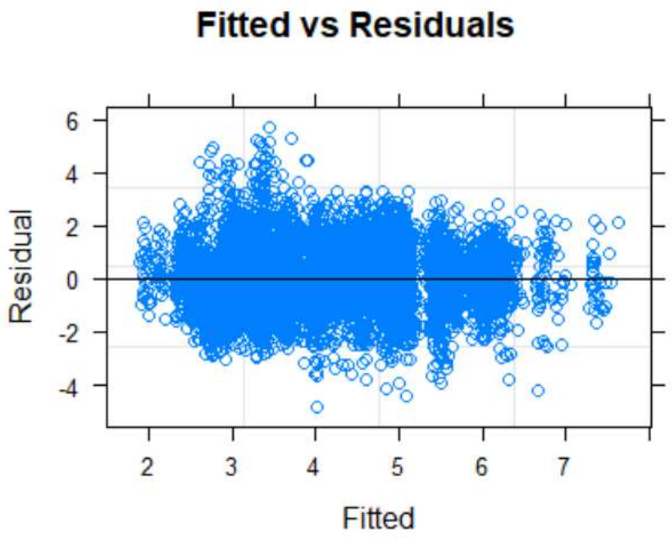
RMSE: 670.4779

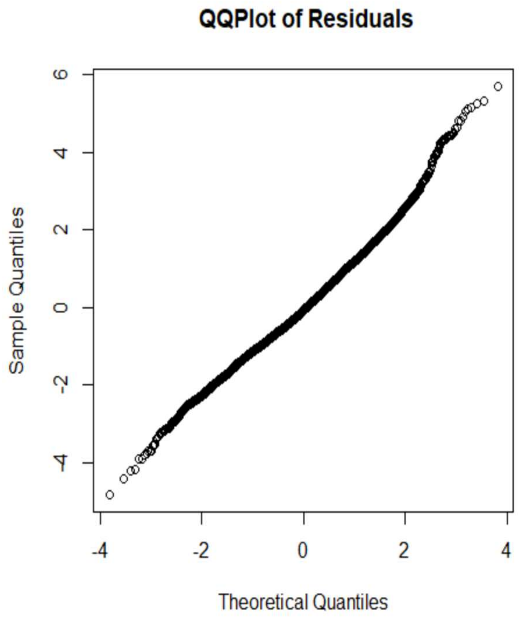


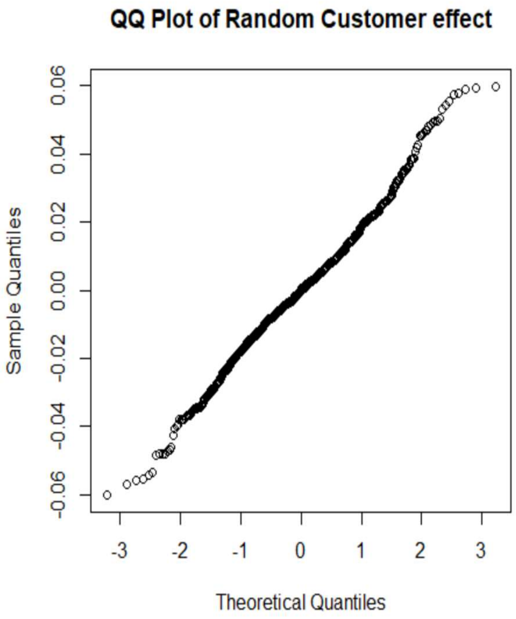
Regression Model Fitting: Mixed Effect

Model: E[Log(Sales)] = State + Sub.Category with random effect : Customer.ID

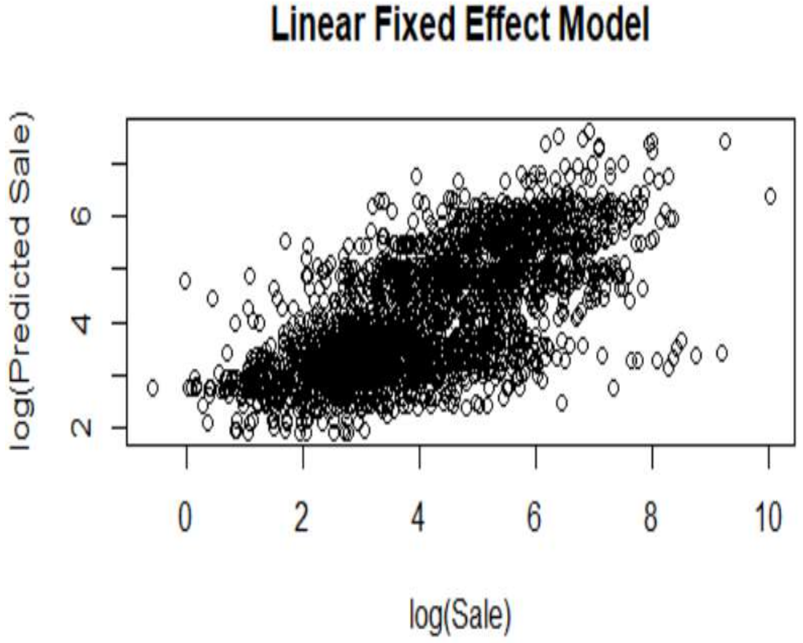
RMSE: 673.1911

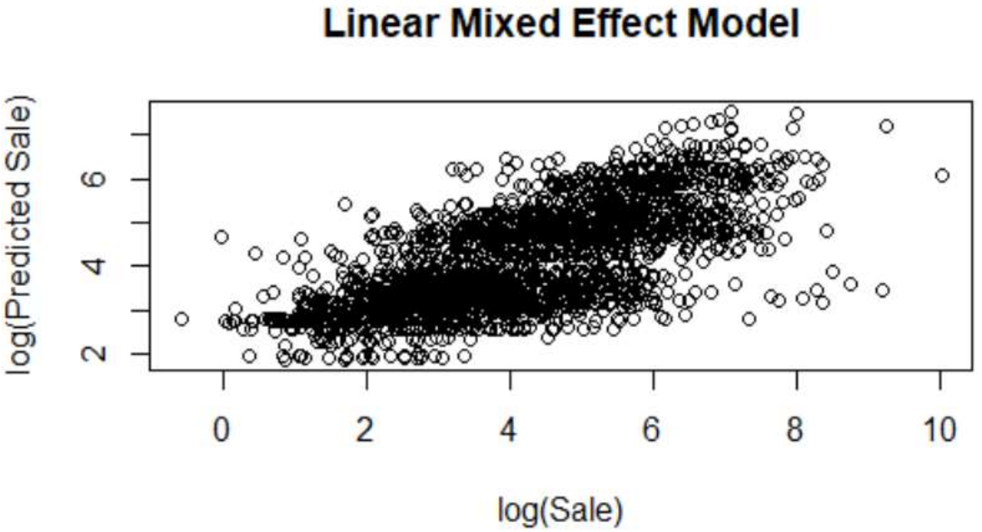

Fitted vs Residuals


QQPlot of Residuals


QQ Plot of Random Customer effect

Regression Model Fitting: Predictive Plots





Regression Model Fitting: LM vs GLM

LM

$$\log(y_i) = \beta_0 + \beta_1 x_1 + \ \ldots + \beta_p x_p + \ \varepsilon_i$$

$$\mu_{\log(y)} = \ \beta_0 + \beta_1 x_1 + \ \ldots + \beta_p x_p$$

GLM

$$\log(\mu_y) = \ \beta_0 + \beta_1 x_1 + \ \ldots + \beta_p x_p$$

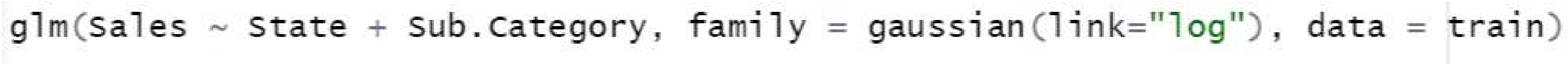
Regression Model Fitting :GLM

GLM with Fixed Effect

RMSE: 781.1244

Model: log[E(Sales)] = State + Sub.Category

Family: Gaussian, Link: Log

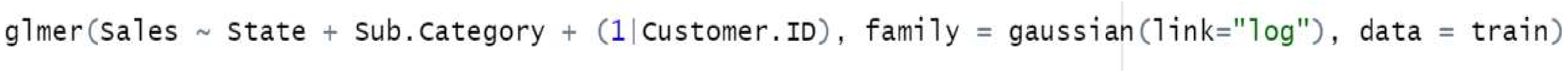


```
glm(Sales ~ State + Sub.Category, family = gaussian(link="log"), data = train)
```

GLM with Random Effect

RMSE: 785.1672

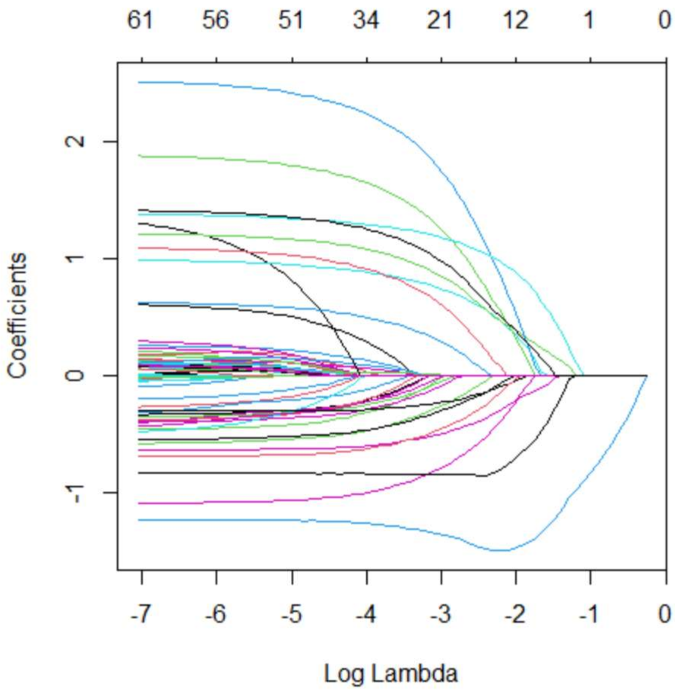
Model: log[E(Sales)] = State + Sub.Category with random effect Customer.ID
Family: Gaussian, Link: Log

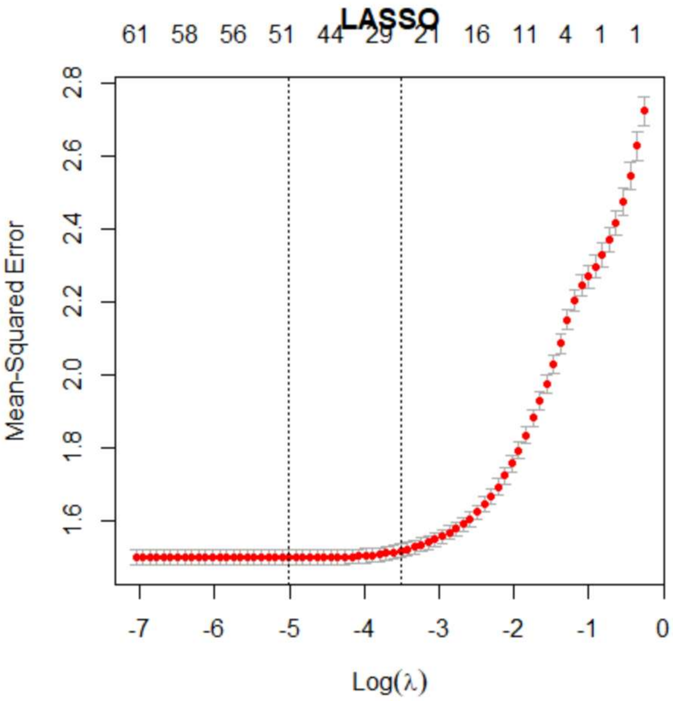


```
glmer(Sales ~ State + Sub.Category + (1|Customer.ID), family = gaussian(link="log"), data = train)
```

Regression Model Fitting : Lasso

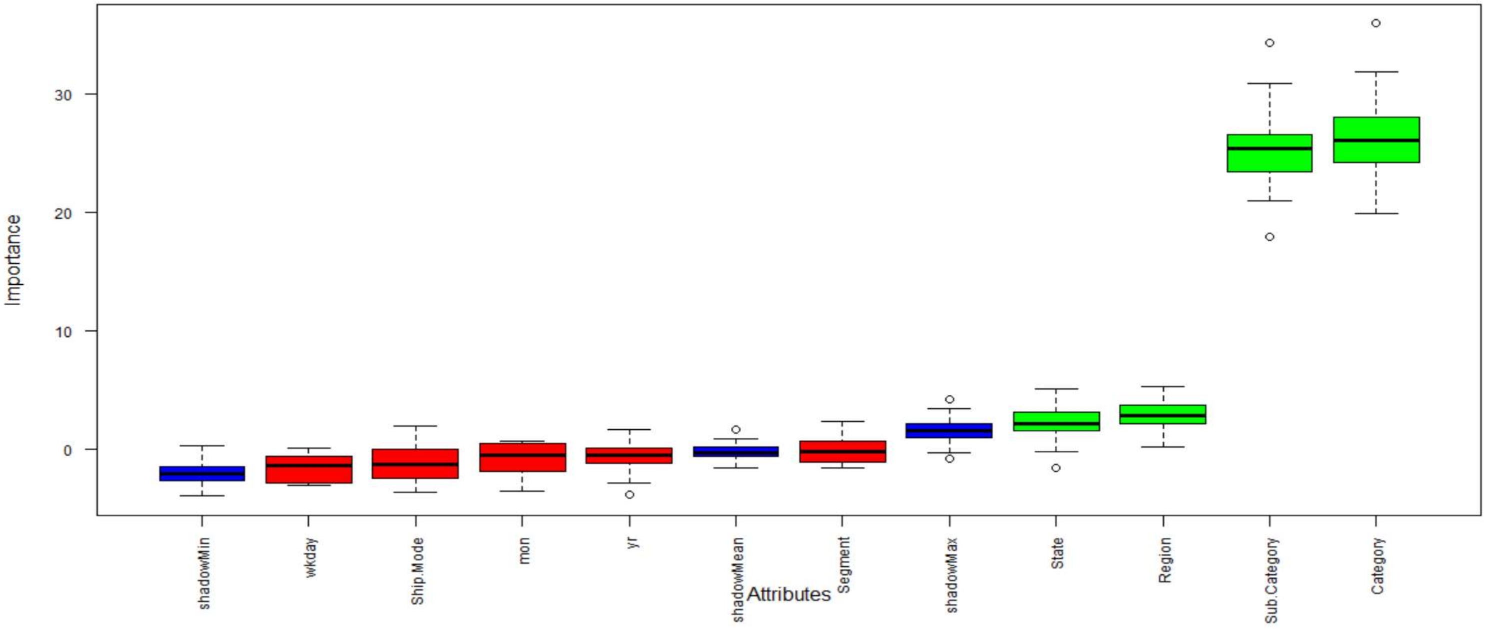
RMSE: 682.3215





Regression Model Fitting: Feature Selection

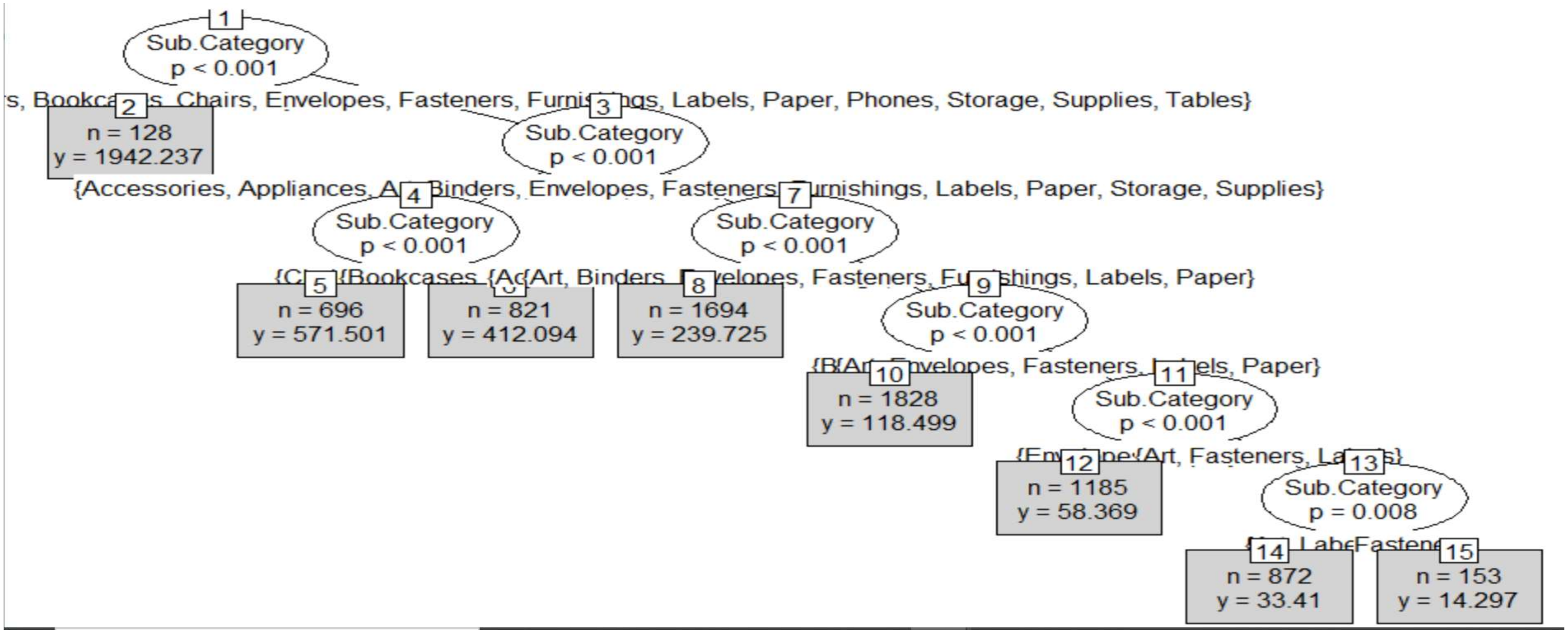
Variance Importance Plot (Boruta)



Regression Model Fitting: Decision Tree

Model: Sales ~ State + Region + Category + Sub.Category

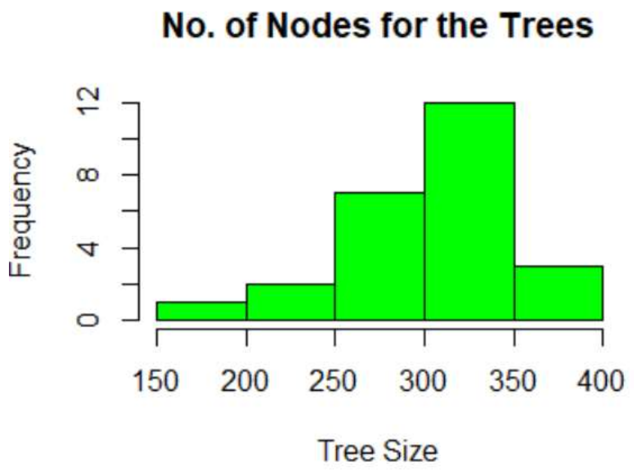
RMSE: 656.501

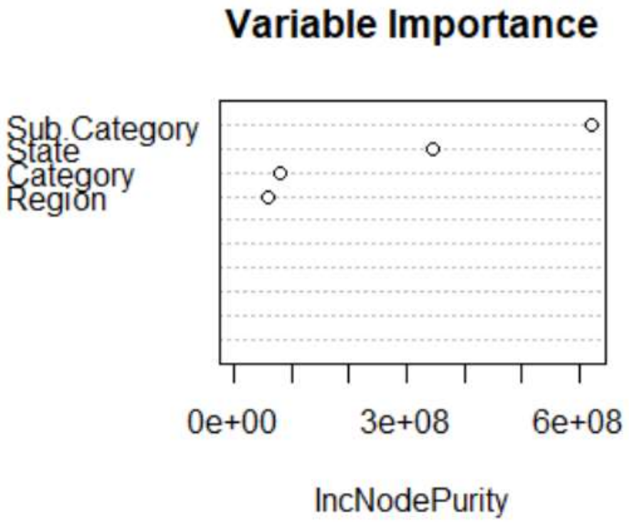


Regression Model Fitting: Random Forest

Model: Sales ~ State + Region + Category + Sub.Category

RMSE: 698.2886

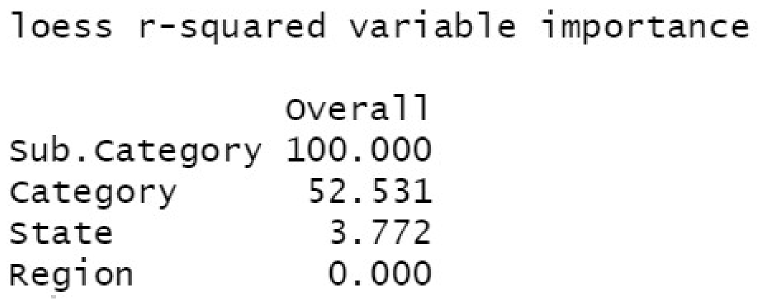




Regression Model Fitting: K Nearest Neighbor

Model: Sales ~ State + Region + Category + Sub.Category

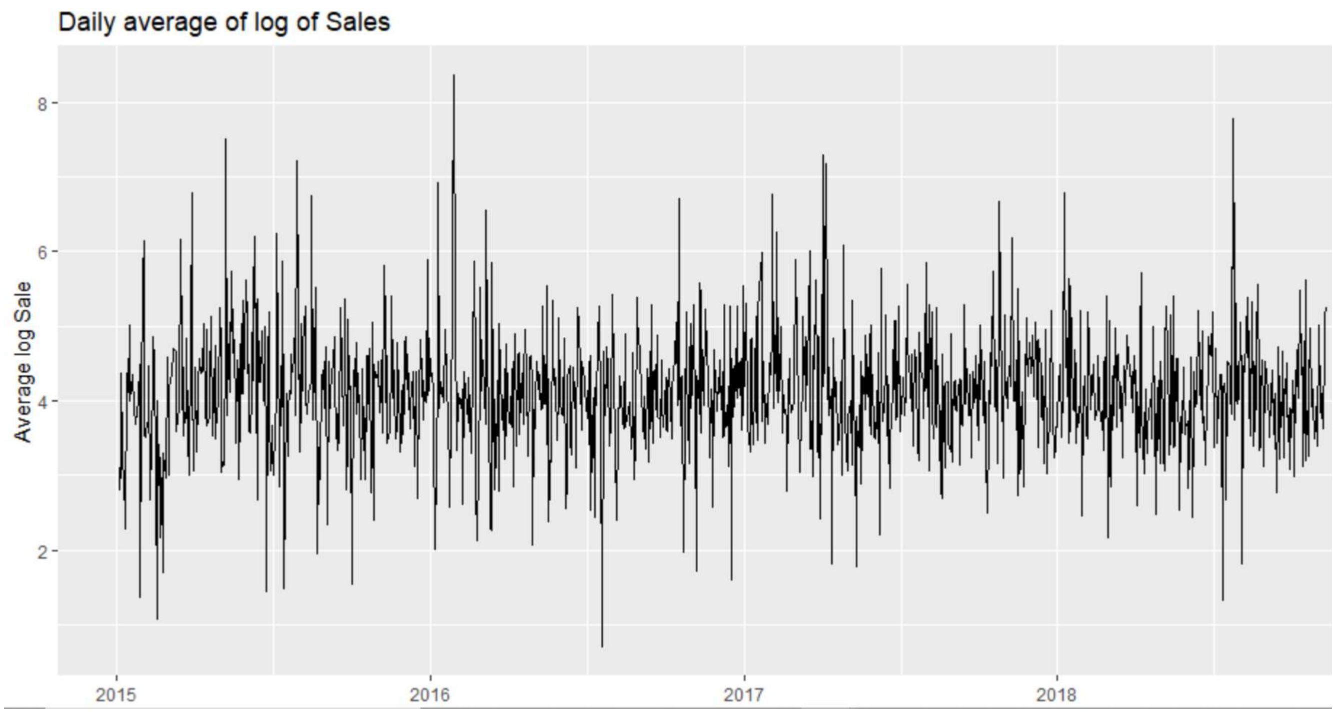
RMSE: 672.4744



```
loess r-squared variable importance

             Overall
Sub.Category 100.000
Category      52.531
State          3.772
Region         0.000
```

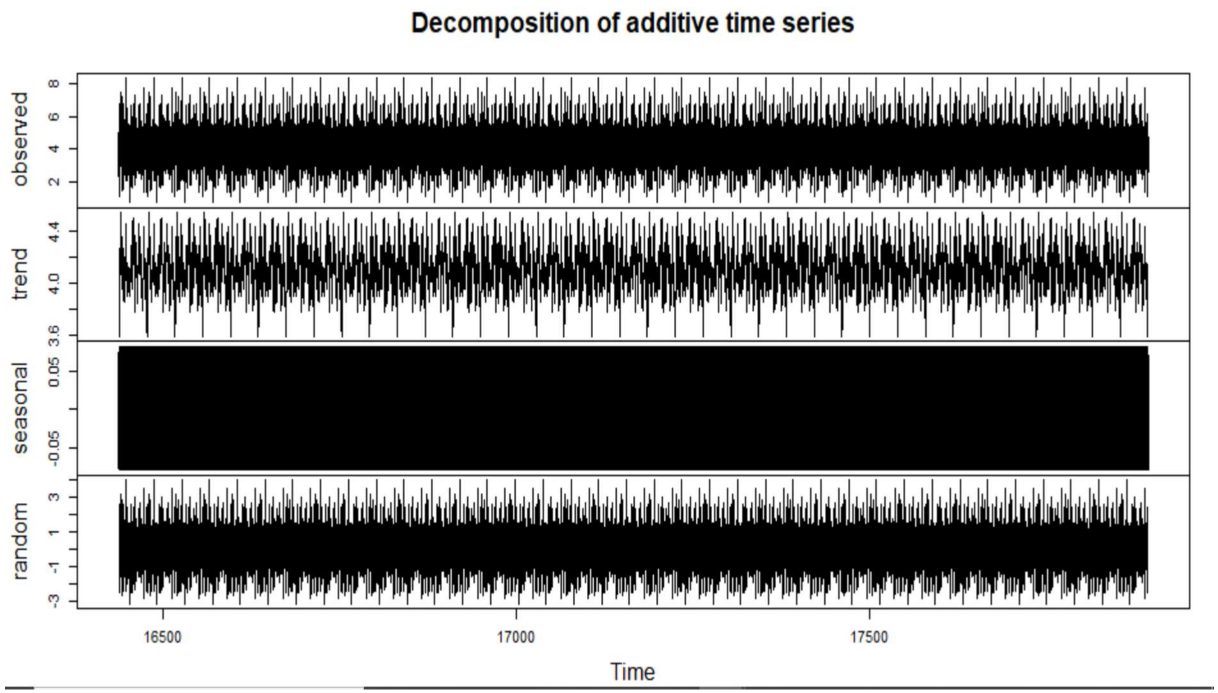
Time Series Analysis



Goal: Predict Sales of 50 days

Response variable: log(sales)

Time Series Analysis



Mann Kendall Trend Test

H_0: no monotonic trend
H_1: trend exists
p-value = 0.93022 > 0.05 (no trend)

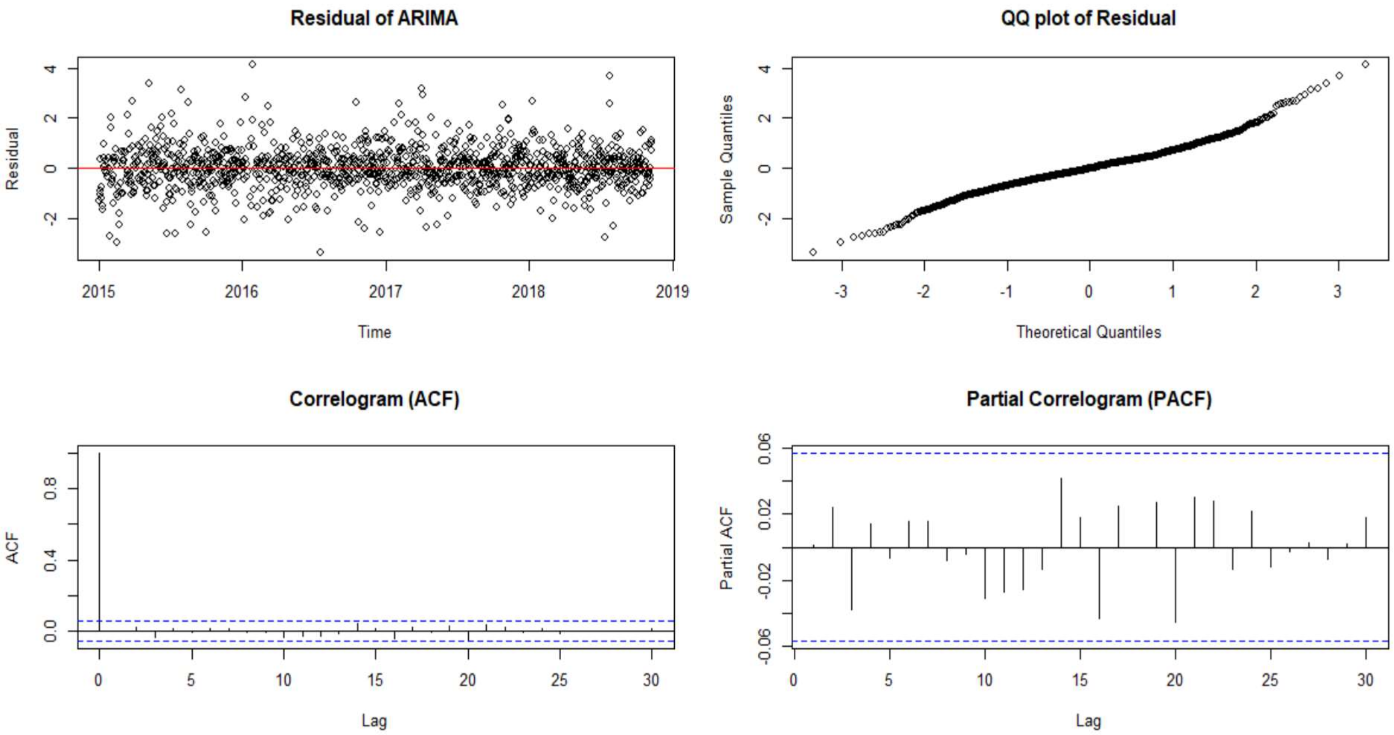
Kruskal-Wallis rank sum test

H_0: location parameters are same in each group
H_1: they differ in at least one
p-value = 0.2811 > 0.05 (no seasonality)

Additive Model: z_t = **log(**$Sales_t$**) =** $Trend_t + Seasonal_t + Random_t$

Time Series Analysis

ARIMA Model: $z_t = 4.1096 + 0.9464 * z_{t-1} - 0.8899 \; * e_{t-1} - 0.0440 \; * e_{t-2} + \; e_t$



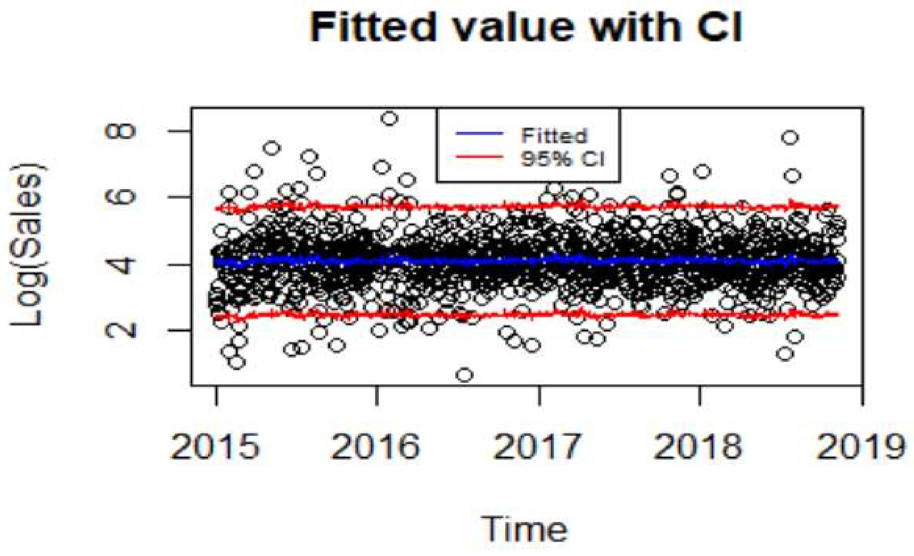
Box-Ljung test
H0: independent
H1: serial correlation

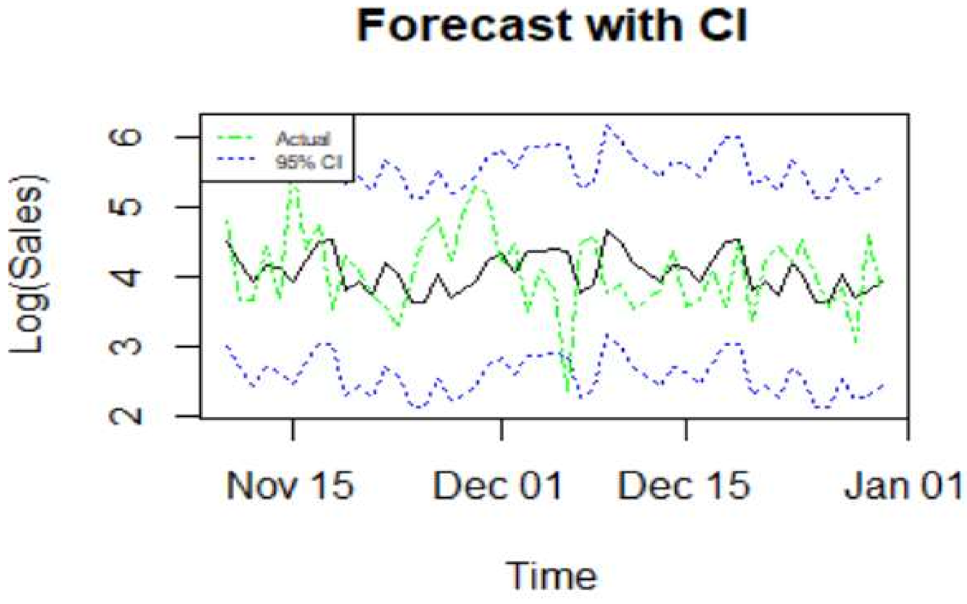
p-value = 0.7717 > 0.05

No need for ARCH or GARCH Model

Time Series Analysis

RMSE: 52.87859

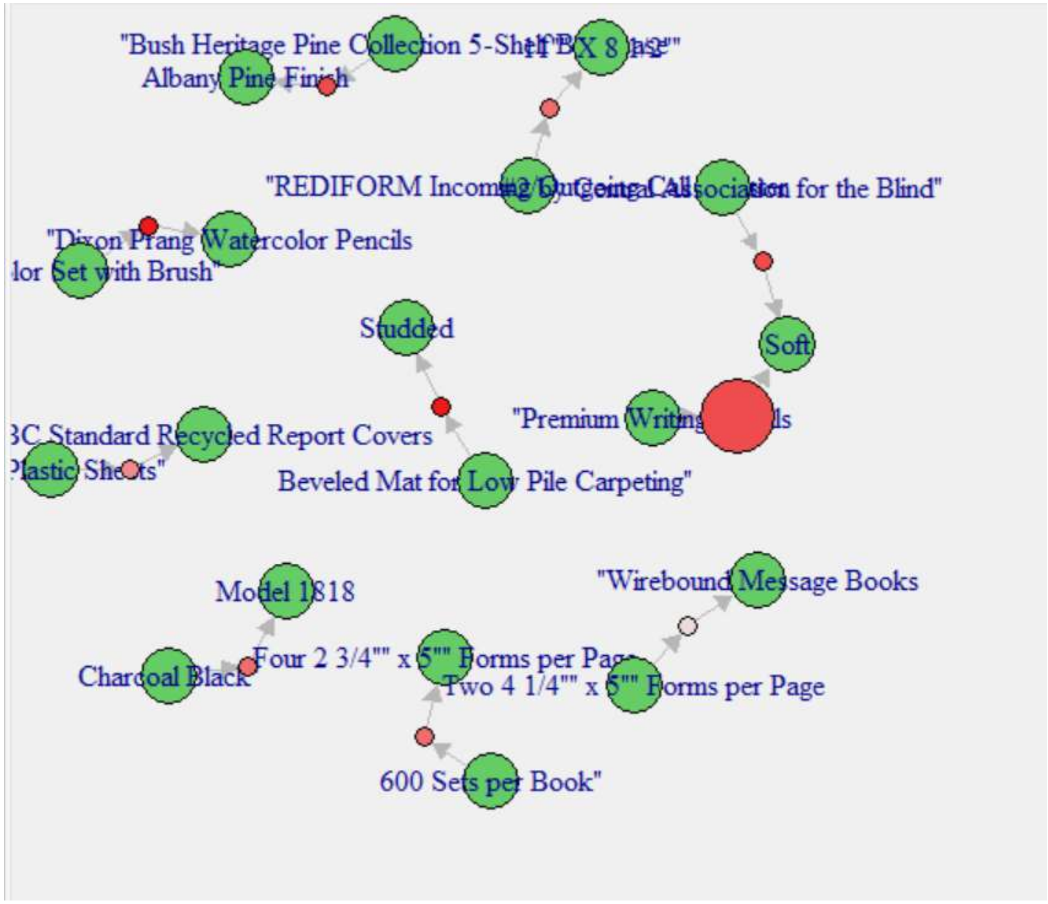




Market Basket Analysis

Top 10 rules by **Confidence**

If A => B is the rule, confidence shows the proportion of transactions having both A and B, out of total transactions having A.

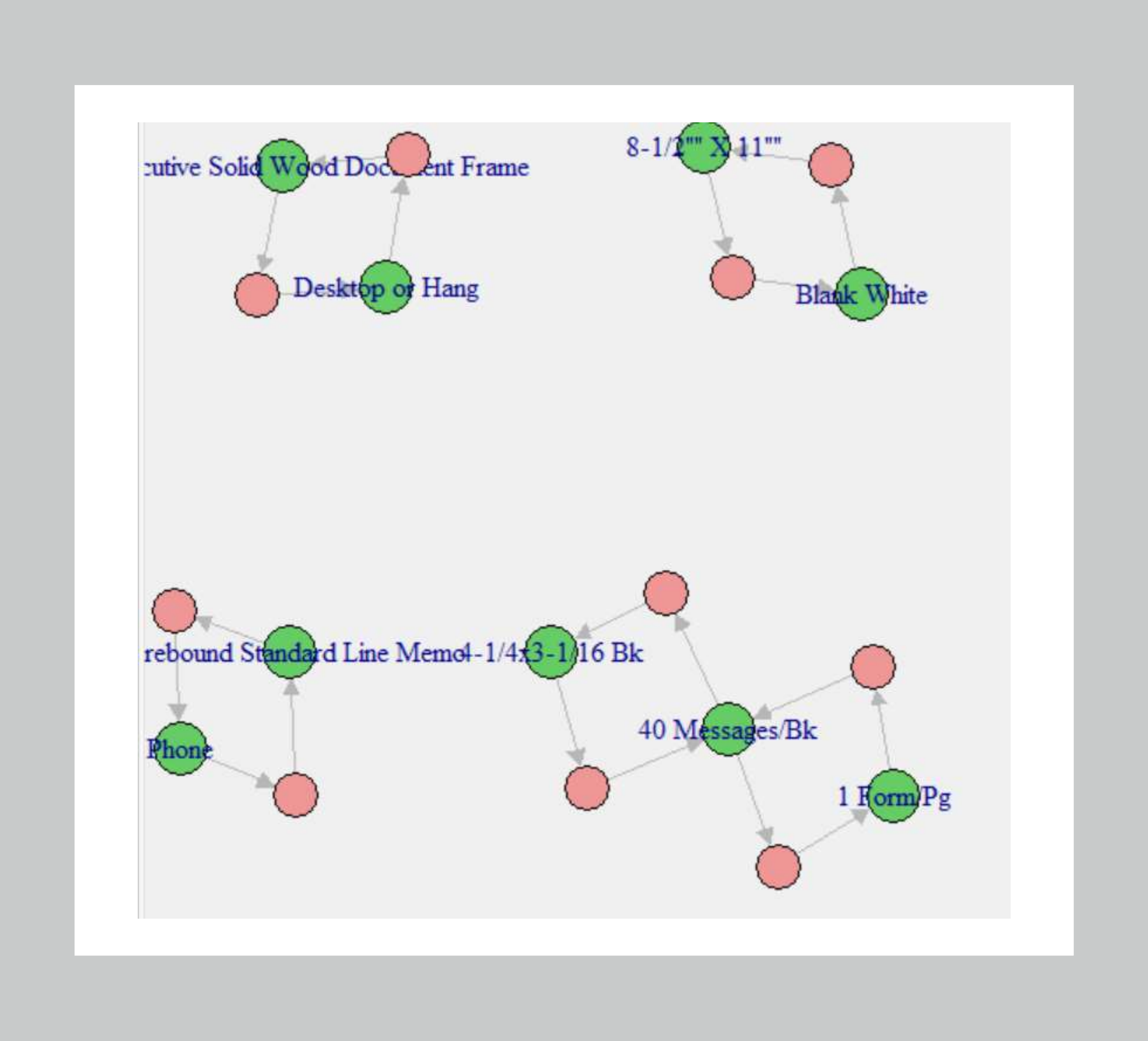

"Bush Heritage Pine Collection 5-Shelf B
11" X 8 1/2"
Albany Pine Finish
"REDIFORM Incoming/Outgoing Call
#2 by General Association for the Blind"
"Dixon Prang Watercolor Pencils
lor Set with Brush"
Studded
Soft
"Premium Writing
ls
3C Standard Recycled Report Covers
Plastic Sheets"
Beveled Mat for Low Pile Carpeting"
Model 1818
"Wirebound Message Books
Charcoal Black
Four 2 3/4"" x 5"" Forms per Page
Two 4 1/4"" x 5"" Forms per Page
600 Sets per Book"

Market Basket Analysis

Top 5 rules by **Lift**

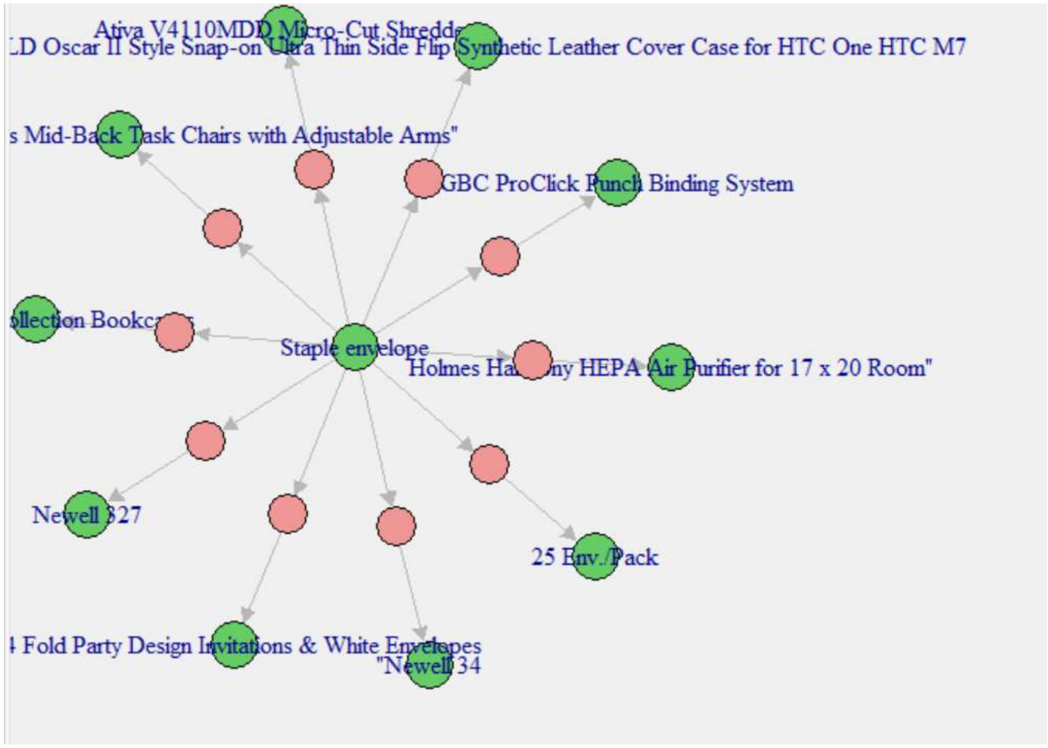
Lift is the factor by which, the joint occurrence of A and B exceeds the expected probability of A and B joint occurring, had they been independent.

Higher the lift, higher the chance of A and B occurring together.



Market Basket Analysis

Staple envelope is the most popular item, we are interested in the items bought with it.


Ativa V4110MDD Micro-Cut Shredder
LD Oscar II Style Snap-on Ultra Thin Side Flip Synthetic Leather Cover Case for HTC One HTC M7
s Mid-Back Task Chairs with Adjustable Arms"
GBC ProClick Punch Binding System
ollection Bookcases
Staple envelope
Holmes Harmony HEPA Air Purifier for 17 x 20 Room"
Newell 327
25 Env./Pack
4 Fold Party Design Invitations & White Envelopes
"Newell 34

Future Scope

Panel Data analysis

Unbalanced Time Series analysis with exploratory variables

Random Forest with Mixed effect model

Bayesian Approach

Thank you!