Forecast Cami's Bakery Sales

Gloria P. Moore

DSC680_Applied Data Science

Bellevue University

Bellevue University Data Science Master's Program

Abstract

The goal of this project is to predict a Home-Based Bakery, using data provided by the business I aim to predict net sales based on the information collected from previous sales and website traffic behavior, it is also within the scope of this project to use graph analysis to draw a data driven marketing strategy. Next sections will take you through the process of collecting, cleaning, formatting, organizing, and plotting the data, to finally propose three models to predict sales, Linear Regression, K Nearest Neighbor and ARIMA, from these models the results of root mean squared error are as follow, ARIMA result a RMSE of 0.89, Linear regression model 0.064, and KNN result on a RMSE 0.88. The deployment of any of these models would be decision of the Bakery owners, predictions for the three models are presented in this paper.

Introduction

Forecasting sales is a commonly used practice in every organization, from these estimates almost every organization can build cost budget, overhead count budget, etc. Practice of budgeting is familiar even in the household, small business work with budgets as well, but this task is performed usually based on past data, making some estimations and with a little bit of intuition we predict some numbers that can be happening in our close future bills or earnings. Taking this action of budgeting and forecasting, it would be useful for small businesses to have a better and accurate tool to make this forecasting and from there build budgets and make other important decisions based on past data. Machine learning is defined by Wikipedia as the study of computer algorithms, these algorithms can be used to build a model based on sample data, known as "training data", to make predictions or decisions without being explicitly programmed to do so [1]. The goal of this project is to apply predictive algorithms to forecast sales on a home-based bakery business, for this goal, there are three popular predictive algorithms, autoregressive integrated moving average (ARIMA), Linear Regression and KNN.

ARIMA is an autoregressive model, what this means is the predictions are based on relations withing the same time series data [2] while Linear regression based its prediction on independent variables in the dataset [3], KNN for this project was trained using datetime as well as independent variables, based on this fact, the KNN algorithm performance is compared with the Linear regression performance.

Method

Data was collected through Square and Wix business' account, both files with the data were in csv format, using Jupyter Notebook and Pandas Library, to read and process the files, files were downloaded by year 2020 and 2021 separately. The following is the description of the process to get the final data set used for the models.

Columns on the Sales data set gotten from Square invoice system contains the following columns:

Date	object	Card Brand	object
Time	object	PAN Suffix	object
Time Zone	object	Device Name	object
Gross Sales	object	Staff Name	object
Discounts	object	Staff ID	float64
Net Sales	object	Details	object
Gift Card Sales	object	Description	object
Tax	object	Event Type	object
Tip	object	Location	object
Partial Refunds	object	Dining Option	float64
Total Collected	object	Customer ID	object
Source	object	Customer Name	object
Card	object	Customer Reference ID	object
Card Entry Methods	object	Device Nickname	float64
Cash	object	Deposit ID	object
Square Gift Card	object	Deposit Date	object
Other Tender	object	Deposit Details	object
Other Tender Type	object	Fee Percentage Rate	object
Other Tender Note	object	Fee Fixed Rate	object
Fees	object	Refund Reason	object
Net Total	object	Discount Name	object
Transaction ID	object	Transaction Status	object
Payment ID	object	Order Reference ID	float64

For each of the Sales file columns with NaN were dropped, based on the fact that columns with NaN were not critical for the analysis:

Card Entry Methods	29	Description	1
Other Tender Type	170	Dining Option	173
Other Tender Note	171	Customer ID	6
Card Brand	29	Customer Name	7
PAN Suffix	29	Customer Reference ID	159
Device Name	120	Device Nickname	173
Staff Name	52	Deposit ID	29
Staff ID	173	Deposit Date	29

```
Deposit Details          29          Refund Reason           172
Fee Percentage Rate      29          Discount Name           171
Fee Fixed Rate           29          Order Reference ID      173
```

Data frame from 2021 year was append using `.append` method to data frame 2020. The complete data set contains sales for both years available, columns of Date and Time was converted to `to_datetime`

For website traffic data, csv files were downloaded from Wix.com these files were organized in years 2020 and 2021, both files contained the following columns:

```
Page_Views                int64
Site_Sessions             int64
Unique_Visitor            int64
Site_Bounce_Rate         object
Session_minutes           int32
Session_seconds          object
```

No missing values were detected, data set corresponding to year 2021 was appended to year 2020, date was transformed to datetime format. Session_minutes column was transformed to seconds and add it to the column Session_seconds. For details of traffic and sales data frame please consult Appendix 1. Sales dataframe and traffic data frame was merged using “inner” method and using “Date” as key variable.

Outilers:

For outliers, the method used was to drop the observations on Sale data frame that represent outliers, for Net sales the observation $1216 was a punctual sale made in December 2020 to a High School in the area, for Session Duration greater than 800 seconds also was dropped because represented a session way out of average time spend on the website.

Box Plot Net sales

Box Plot Session_Duration

Final Data frame

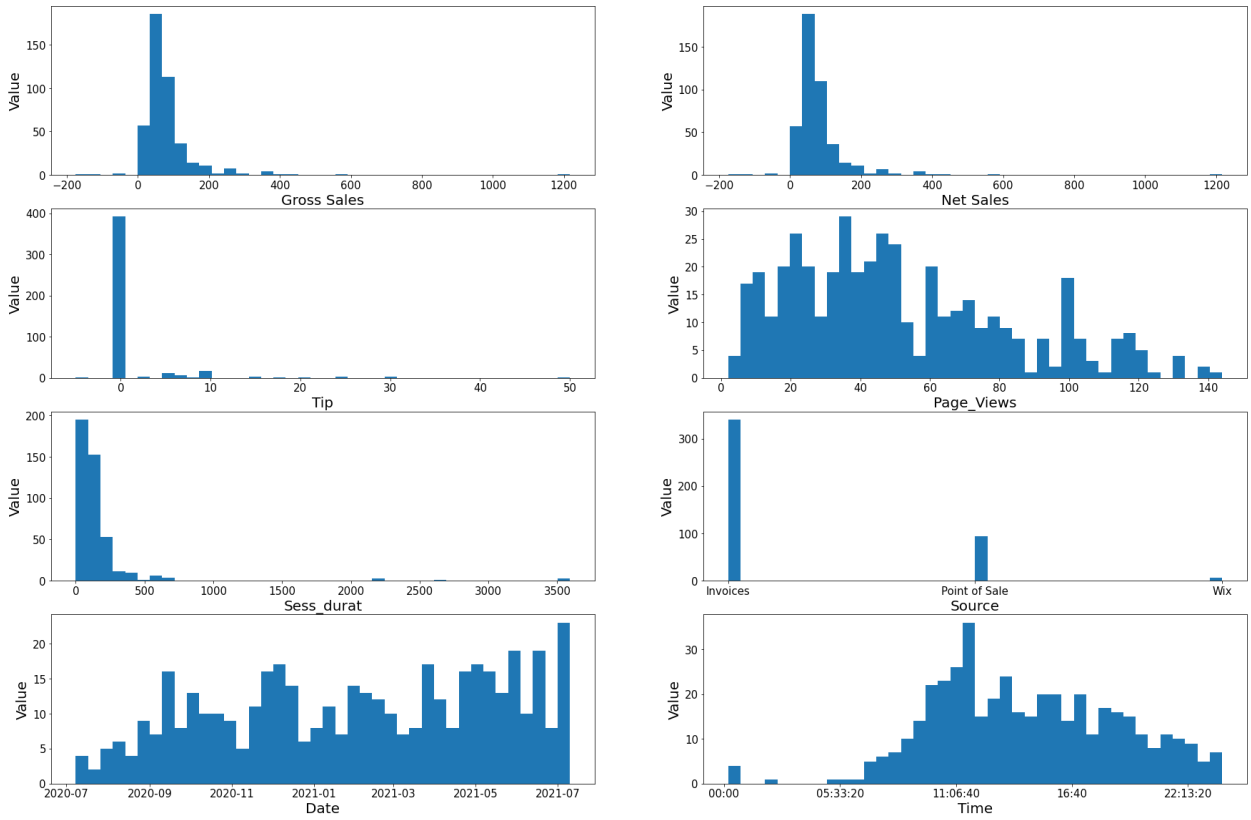
[21]:

	Date	Time	Gross Sales	Net Sales	Tip	Source	Page_Views	Site_Sessions	Unique_Visitor	Site_Bounce_Rate	Sess_durat
0	2020-12-31	14:02:13	55	55	0.0	Invoices	11	6	6	67.00%	32
1	2020-12-28	08:56:42	45	45	0.0	Invoices	30	9	9	44.00%	235
2	2020-12-27	10:17:39	130	130	0.0	Point of Sale	15	7	7	57.00%	7
3	2020-12-26	17:06:22	60	60	0.0	Invoices	30	12	10	50.00%	89
4	2020-12-23	13:54:09	45	45	0.0	Invoices	19	10	10	70.00%	20
...	...	...	...	...	...	...	...	...	...	...	...
435	2021-01-07	09:06:12	150.00	150.00	0.0	Invoices	23	10	9	60.00%	93
436	2021-01-06	11:42:21	55.00	55.00	0.0	Invoices	26	10	10	40.00%	46
437	2021-01-05	12:58:16	360.00	360.00	0.0	Point of Sale	17	8	8	75.00%	14
438	2021-01-02	19:07:01	75.00	75.00	0.0	Invoices	32	6	6	17.00%	83
439	2021-01-02	09:46:23	47.00	47.00	0.0	Invoices	32	6	6	17.00%	83

440 rows × 11 columns

Analyzing the target variable and other features in order support marketing strategies, different plots were created.

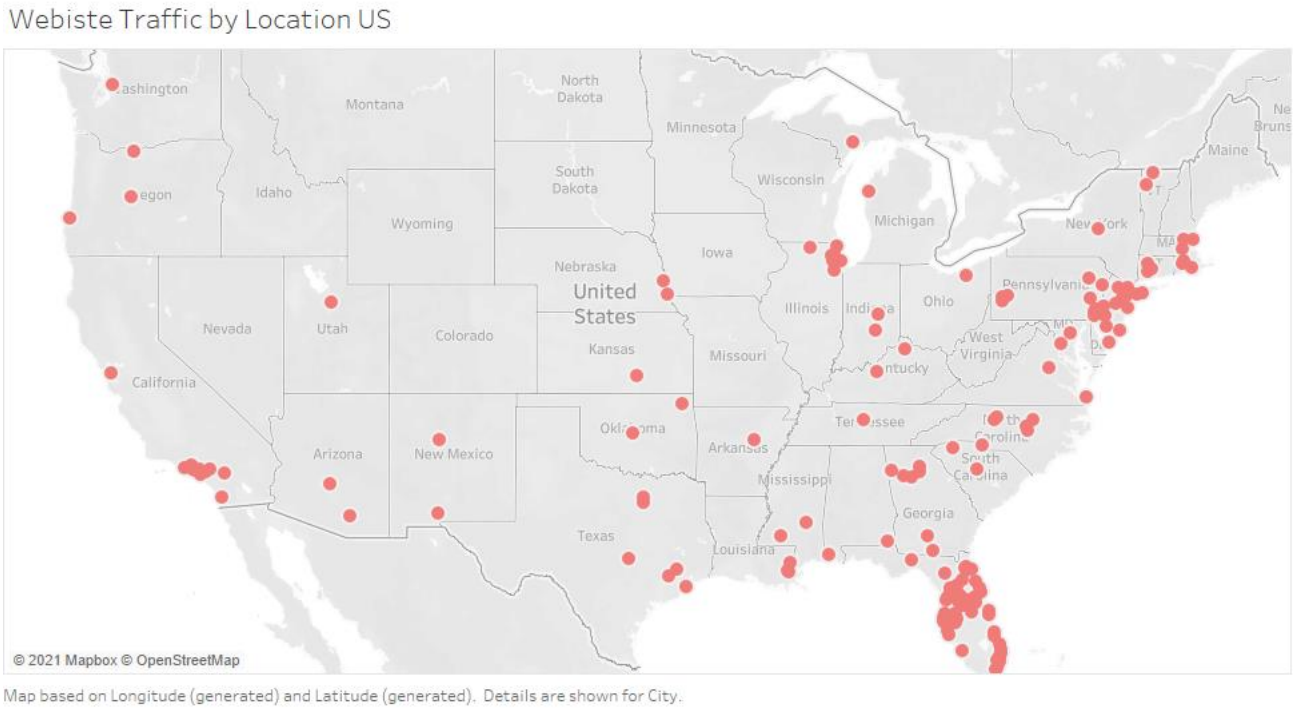
Histograms



These Histograms are used to support marketing strategies like:

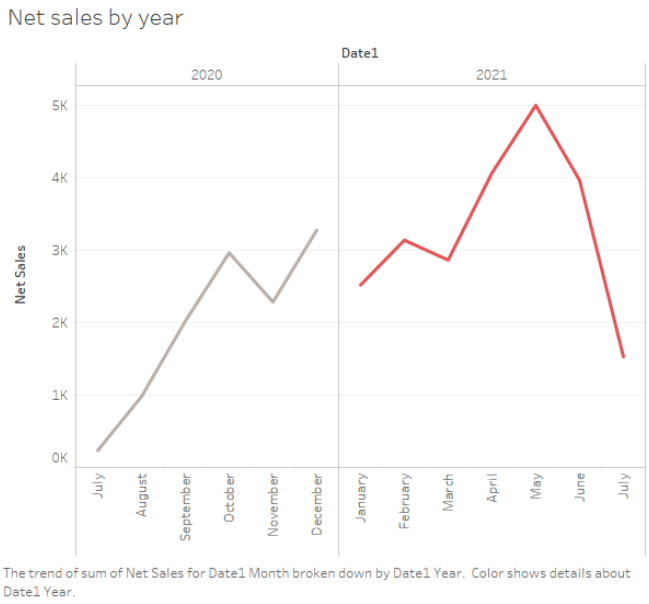
- The ends of each month present high volume of sales, what brings the idea to increase promotions and Sales strategies for the second week of each month to increase sales.
- Most of the sales have an average value of $100, this will be helpful to design wedding cake sale strategies, wedding cakes are more expensive what will bring Net sales up.
- Most of the sales does not generate tips, usually small business does not receive tips, what it might be removed as a request from the invoice system.

Other graph analysis that was made is the website visits by location, this graph was built in Tableau.

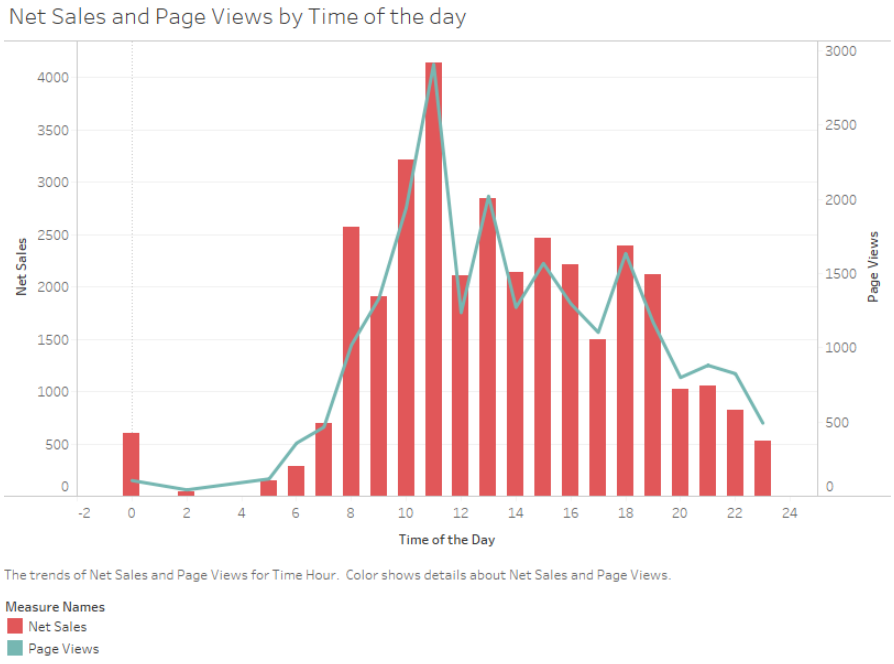


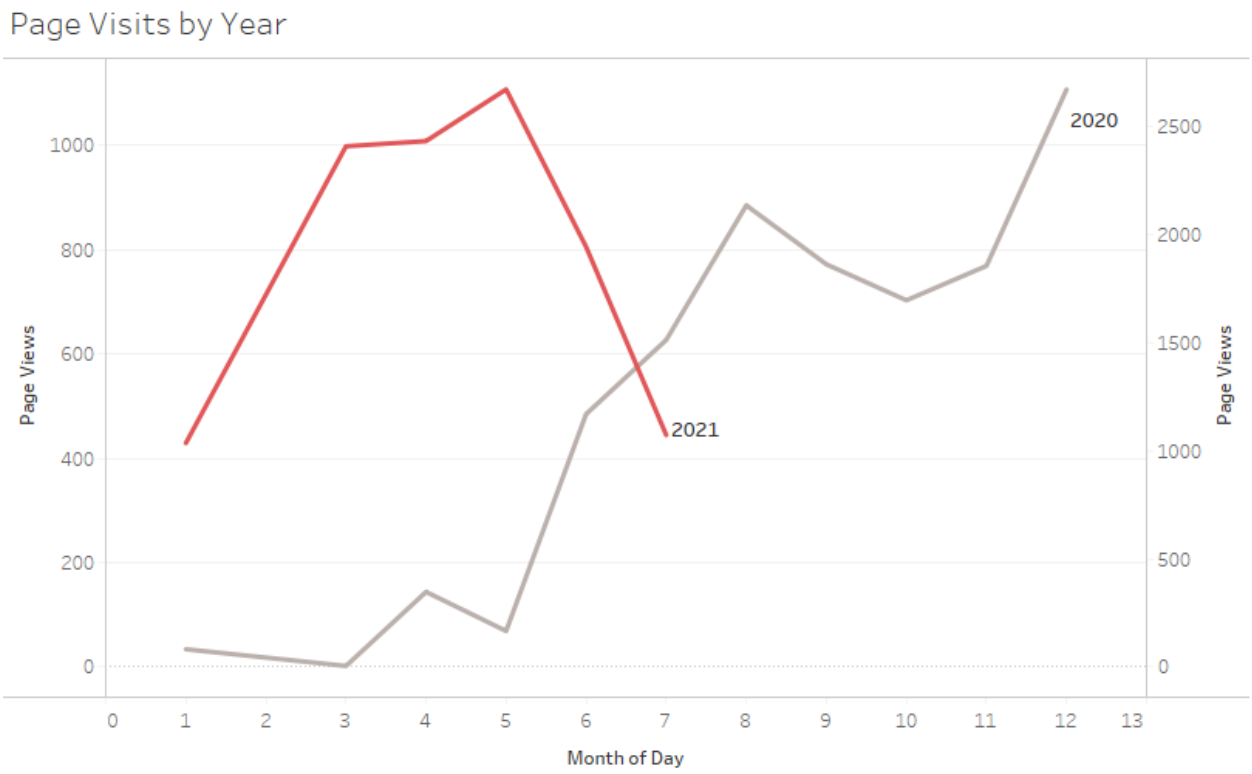
Most of the clients as expected, come from Florida, but it is interesting how many visits I get from the whole East Coast, this might be because the bakery offers delivery, but based on this analysis, we started a new promotion of free delivery to Disney with orders of $100 or more. Future analysis will determine how effective this strategy went. We are starting another project to ship cakes out of the State, based on the number of visitors from all over the country.

The target variable was defined as Net sales, Gross sales was dropped because it did not have any variation compared with Net Sales, the graph analysis for the target variable include Line graph to compare sales in 2020 and 2021 by month:



Sales was compared with Page Visits by Time of the Day



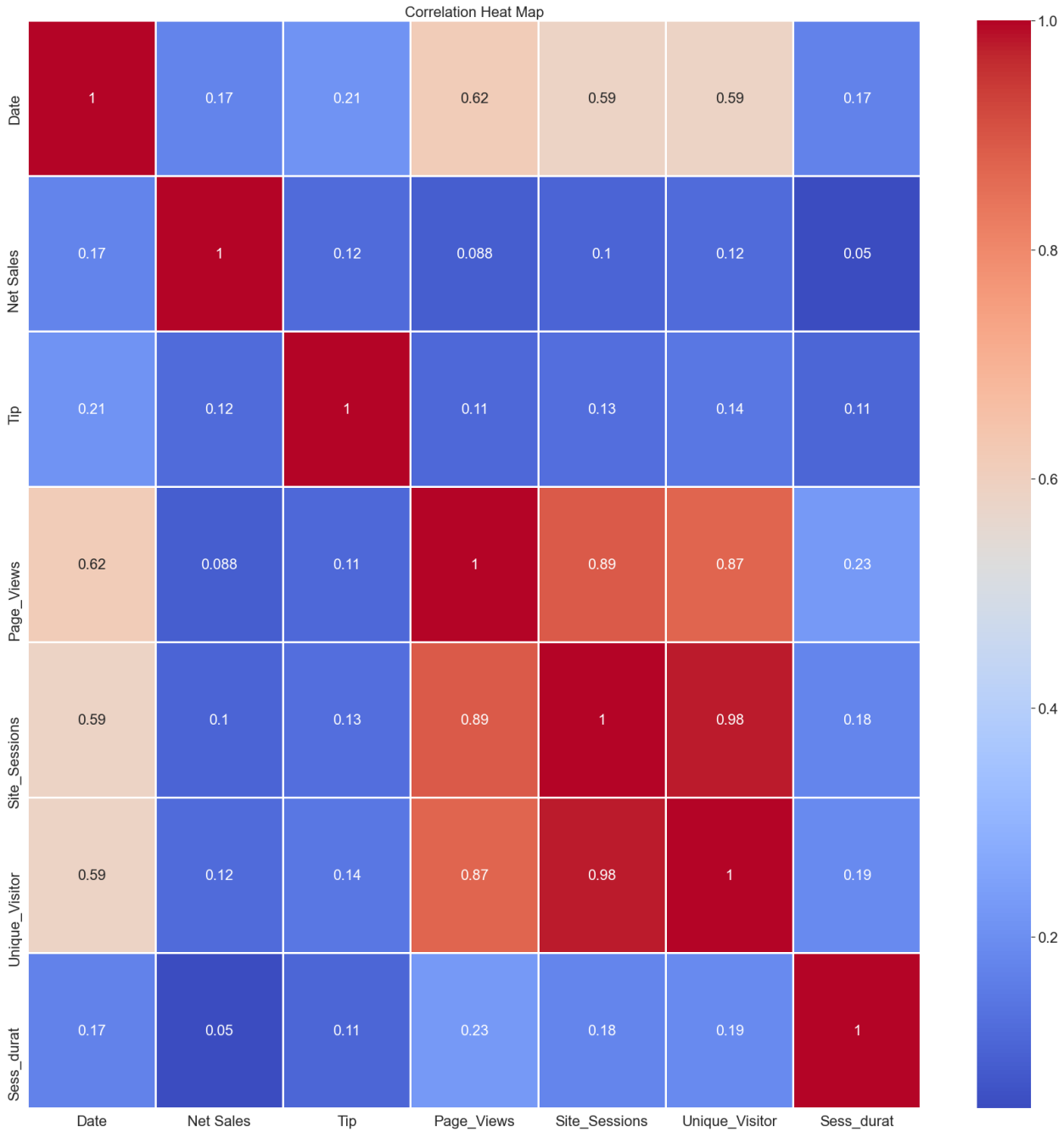


From the previous information, the following marketing projects has been started:

- Social media posts will be done in between 8 am and 2 pm.
- More promotions and new products will be announced at the end of the months for the next month/season.

Modeling

For the modeling stage, 2 different date sets were prepared, the first one was used for linear regression, and the Date column was transformed to and ordinal number using `.toordinal` method, and column Gross Sales and Site_Bounce_rate were dropped. Correlation Heat Map as follow:



From this graph, there Is not high correlation value in between Net sales and the other variables. However, the date is highly correlated with the website traffic variables, for scatterplot matrix please go to Appendix 1.

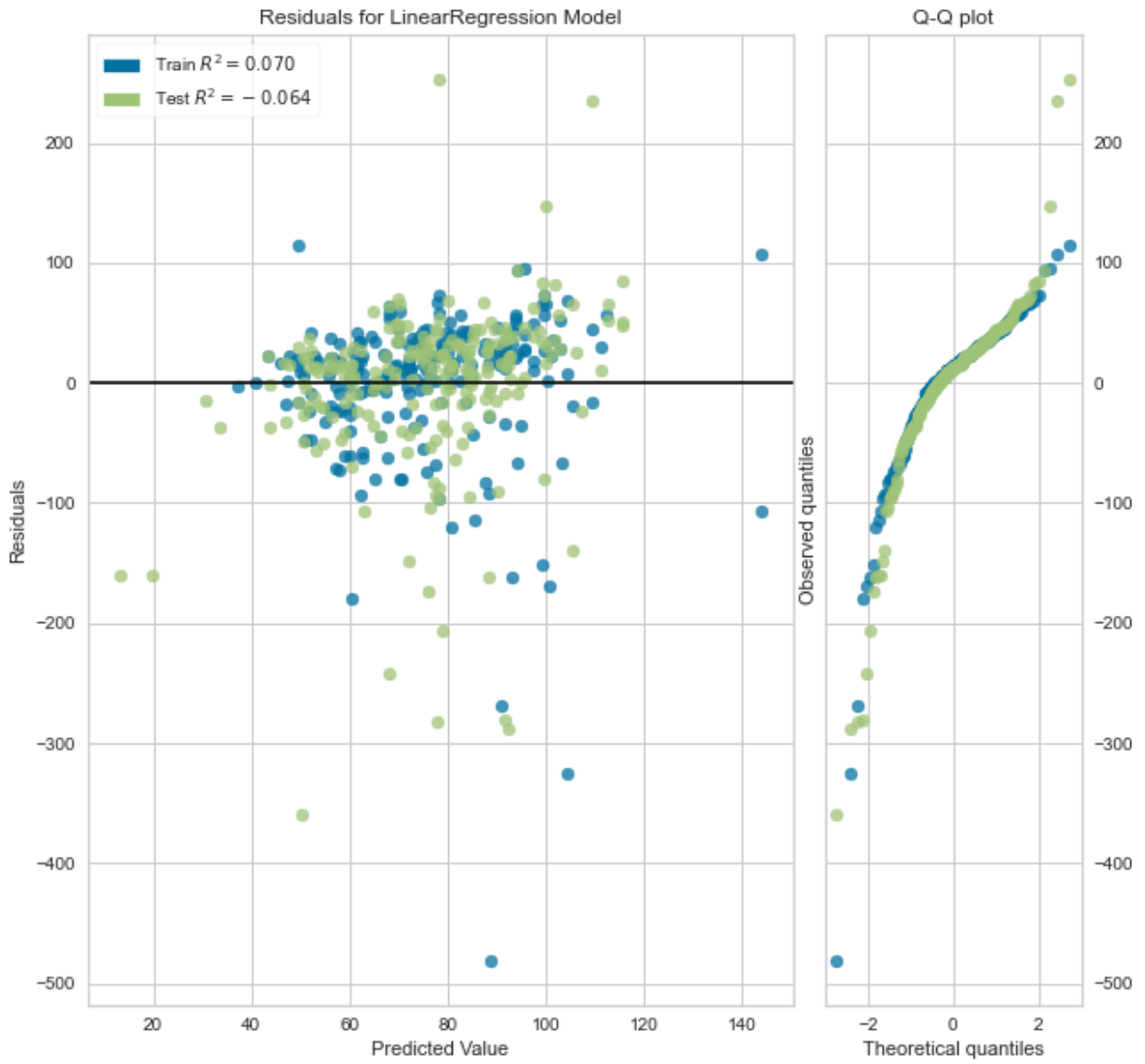
Linear regression

For my first model I used Linear regression, this approach uses independent variables to predict my target variable, using the following features, the model was built:

```
Date
Tip
Page_Views
Site_Sessions
Unique_Visitor
Sess_durat
```

The model Trained have the following results:

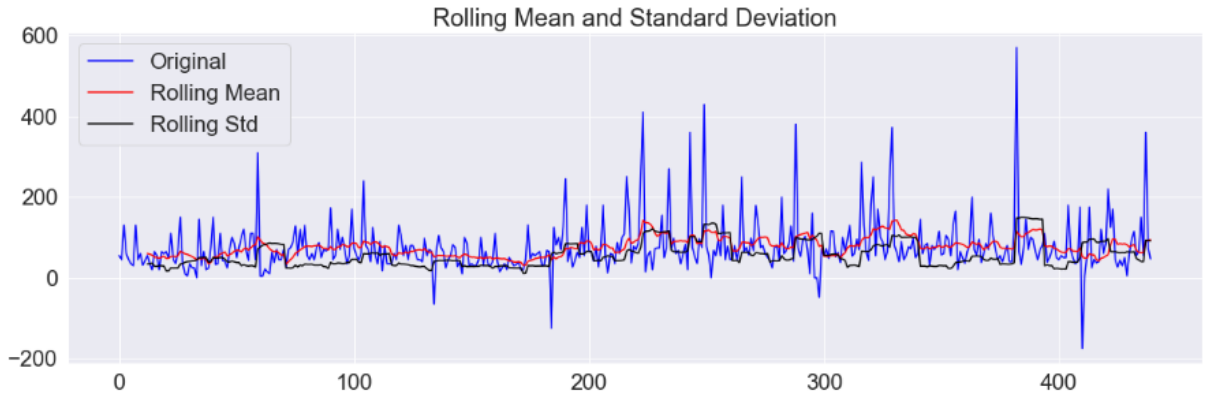
```
Intercept:
 -132878.79609816466
Coefficients:
 [ 0.18021187 -1.62813419 -0.02725015 -3.65368809  3.29653761  0.02312235]
R2:
 0.06957097291236392
```



For the Residuals graphs, RMSE equals to 0.07 for the train data set and 0.064 for the test data set, the residuals look distributed around 0, and the Q-Q plot shows a distribution close to normal on the values that are not to far from the mean. For detailed graph Actual vs Predicted Values please check Appendix 2.

ARIMA:

Arima model was trained using time series analysis, a Dickey Fuller test was applied to the target variable Net Sales to verify stationarity of it:



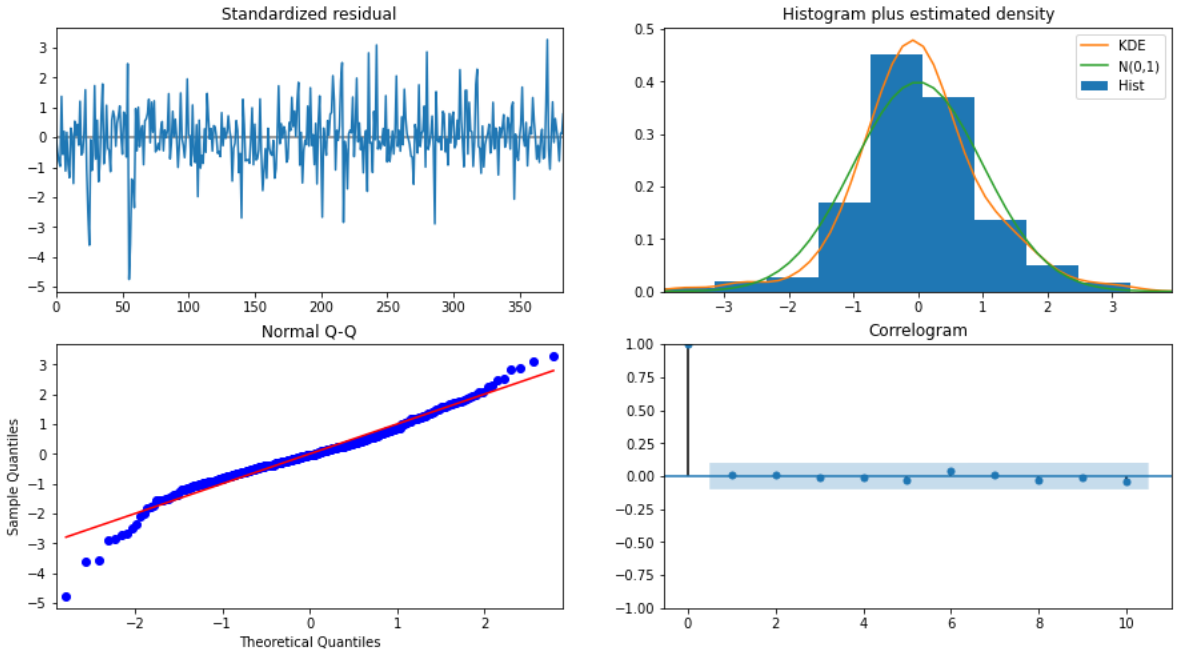

```
Results of dickey fuller test
Test Statistics                -19.289351
p-value                          0.000000
No. of lags used                 0.000000
Number of observations used    438.000000
critical value (1%)             -3.445368
critical value (5%)             -2.868161
critical value (10%)            -2.570297
dtype: float64
```

The Nule hypothesis of the Dickey Fuller test is The series has a unit root (value of a =1) [4], with a p-value 0.00, and with a test statistic smaller than all critical values, we can reject the Null Hipthesys and assume that the target variable is stationary, after this test the ARIMA model can be used without any data processing more than that what was already done. For this test a the target variable was used in the units that was available (Apendix 3), and after this attempt I applied `.log` to the variable and train the model using this variable, these were the resutls

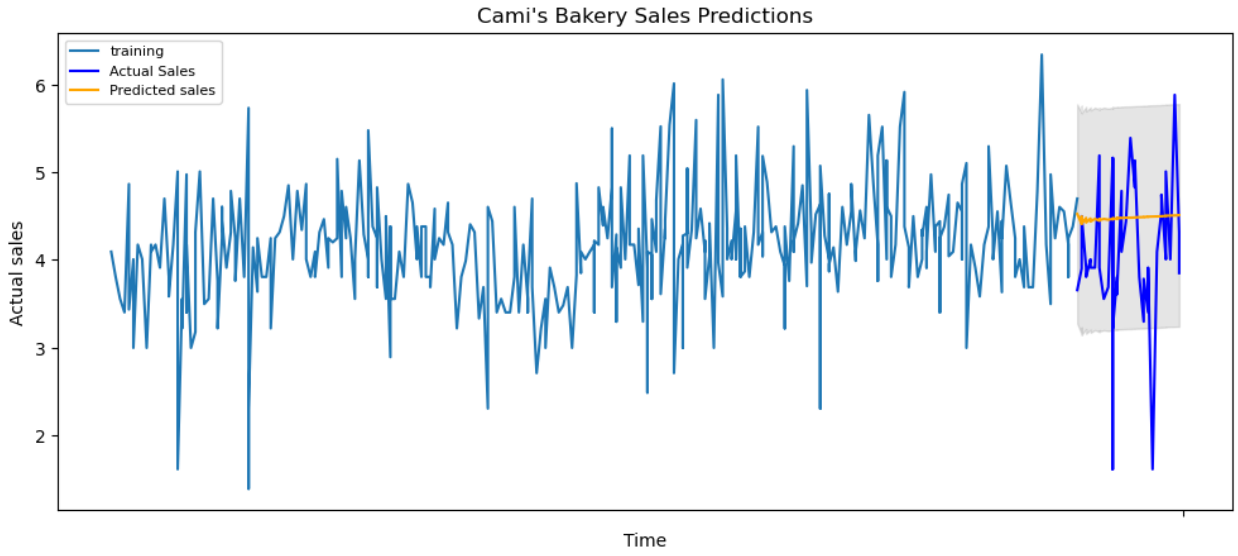
Applying autoARIMA to determine the best order for this dataset [5]:

```
Performing stepwise search to minimize aic
 ARIMA(0,0,0)(0,0,0)[0]             : AIC=2193.227, Time=0.01 sec
 ARIMA(1,0,0)(0,0,0)[0]             : AIC=970.181, Time=0.03 sec
 ARIMA(0,0,1)(0,0,0)[0]             : AIC=1781.959, Time=0.04 sec
 ARIMA(2,0,0)(0,0,0)[0]             : AIC=inf, Time=0.05 sec
 ARIMA(1,0,1)(0,0,0)[0]             : AIC=772.701, Time=0.14 sec
 ARIMA(2,0,1)(0,0,0)[0]             : AIC=770.053, Time=0.23 sec
 ARIMA(3,0,1)(0,0,0)[0]             : AIC=771.969, Time=0.30 sec
 ARIMA(2,0,2)(0,0,0)[0]             : AIC=770.119, Time=0.62 sec
 ARIMA(1,0,2)(0,0,0)[0]             : AIC=769.886, Time=0.26 sec
 ARIMA(0,0,2)(0,0,0)[0]             : AIC=1567.115, Time=0.08 sec
 ARIMA(1,0,3)(0,0,0)[0]             : AIC=771.761, Time=0.33 sec
 ARIMA(0,0,3)(0,0,0)[0]             : AIC=1410.763, Time=0.11 sec
 ARIMA(2,0,3)(0,0,0)[0]             : AIC=inf, Time=0.37 sec
 ARIMA(1,0,2)(0,0,0)[0] intercept   : AIC=768.564, Time=0.63 sec
 ARIMA(0,0,2)(0,0,0)[0] intercept   : AIC=776.488, Time=0.09 sec
 ARIMA(1,0,1)(0,0,0)[0] intercept   : AIC=768.259, Time=0.41 sec
 ARIMA(0,0,1)(0,0,0)[0] intercept   : AIC=775.091, Time=0.05 sec
 ARIMA(1,0,0)(0,0,0)[0] intercept   : AIC=773.883, Time=0.05 sec
 ARIMA(2,0,1)(0,0,0)[0] intercept   : AIC=770.182, Time=0.48 sec
 ARIMA(0,0,0)(0,0,0)[0] intercept   : AIC=786.442, Time=0.04 sec
 ARIMA(2,0,0)(0,0,0)[0] intercept   : AIC=774.950, Time=0.06 sec
 ARIMA(2,0,2)(0,0,0)[0] intercept   : AIC=764.964, Time=0.70 sec
 ARIMA(3,0,2)(0,0,0)[0] intercept   : AIC=768.079, Time=0.86 sec
 ARIMA(2,0,3)(0,0,0)[0] intercept   : AIC=768.644, Time=0.74 sec
 ARIMA(1,0,3)(0,0,0)[0] intercept   : AIC=769.220, Time=0.71 sec
 ARIMA(3,0,1)(0,0,0)[0] intercept   : AIC=770.677, Time=0.81 sec
 ARIMA(3,0,3)(0,0,0)[0] intercept   : AIC=770.778, Time=0.89 sec

Best model:  ARIMA(2,0,2)(0,0,0)[0] intercept
Total fit time: 9.112 seconds
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  384
Model:               SARIMAX(2, 0, 2)   Log Likelihood                -376.482
Date:                Wed, 21 Jul 2021   AIC                            764.964
Time:                        21:17:45   BIC                            788.668
Sample:                             0   HQIC                           774.366
                                  - 384
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.4406      0.374      1.177      0.239      -0.293       1.174
ar.L1          0.1462      0.146      1.005      0.315      -0.139       0.432
ar.L2          0.7475      0.113      6.616      0.000       0.526       0.969
ma.L1          0.0081      0.141      0.057      0.954      -0.269       0.285
ma.L2         -0.7592      0.089     -8.574      0.000      -0.933      -0.586
sigma2         0.4173      0.022     19.203      0.000       0.375       0.460
===================================================================================
Ljung-Box (Q):                       38.73   Jarque-Bera (JB):                96.31
Prob(Q):                              0.53   Prob(JB):                         0.00
Heteroskedasticity (H):               0.73   Skew:                            -0.25
Prob(H) (two-sided):                  0.07   Kurtosis:                         5.40
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The resultant best model was (2,0,2), training and predicting with the order:



Results Predicted values



Metrics:

```
MSE: 0.7951642582036615
MAE: 0.6933161122427687
RMSE: 0.8917198316756567
MAPE: 0.22042854285562863
```

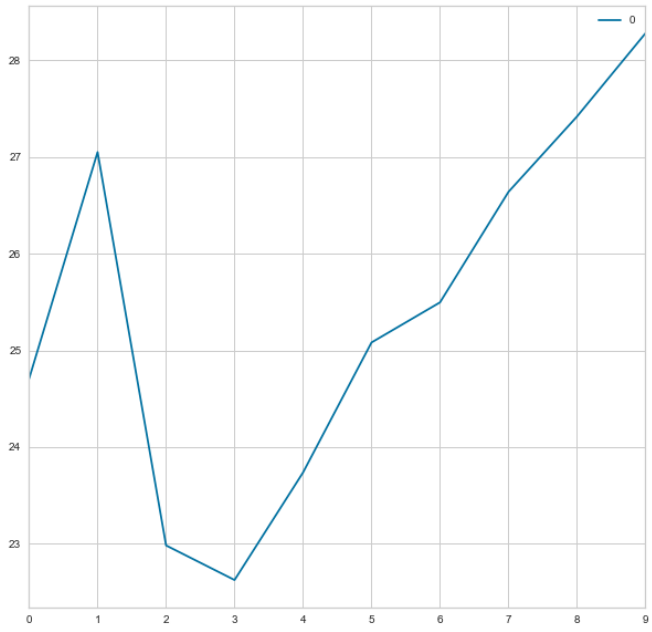
Because this model was built using a logarithmic scale on the target variable, the RMSE resultant is also in the same unit.

KNN

For KNN model the dataset used had the following columns:

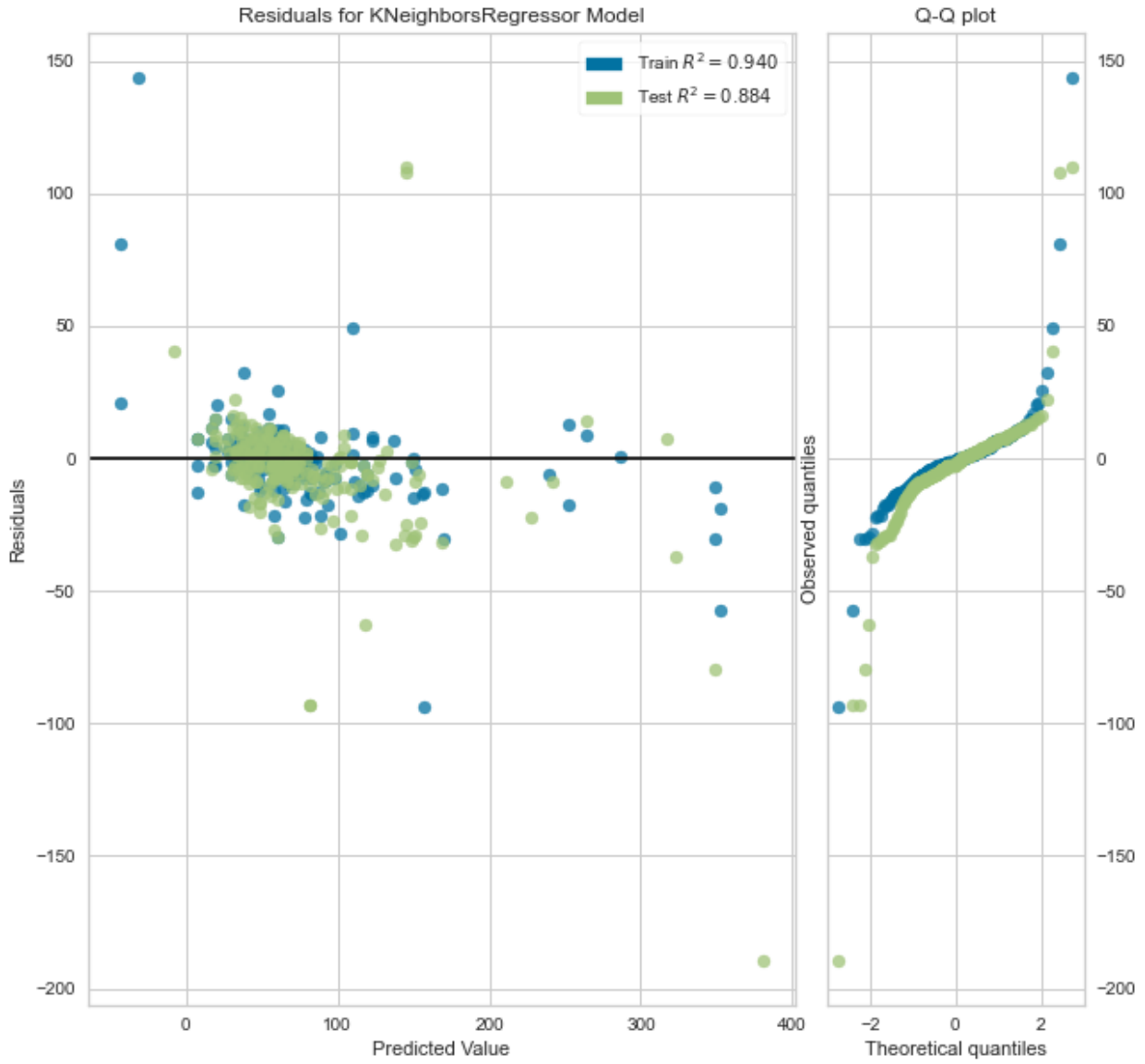
```
Date                0
Time                0
Gross Sales         0
Net Sales           0
Tip                 0
Source              0
Page_Views          0
Site_Sessions       0
Unique_Visitor      0
Site_Bounce_Rate    0
Sess_durat          0
```

All non numerical columns were transformed in dummy variables, and the whole data set was used to predict the Net sales, an elbow graph was created to determine best number for k:




```
RMSE value for k=  1 is: 24.7074
RMSE value for k=  2 is: 27.0520
RMSE value for k=  3 is: 22.9813
RMSE value for k=  4 is: 22.6234
RMSE value for k=  5 is: 23.7360
RMSE value for k=  10 is: 28.2812
RMSE value for k=  6 is: 25.0816
RMSE value for k=  7 is: 25.4964
RMSE value for k=  8 is: 26.6402
RMSE value for k=  9 is: 27.4197
```

Smaller RMSE corresponded to k = 4



For this model the RMSE was 0.88, Q-Q plot for the residuals shows a deviation from the normal from values smaller than the mean, if we compared this plot with the Linear regression plot, these residuals look to be less close to a normal distribution than the Linear Regression residual's, for details of tested values and predicted values please see Appendix 4

Model Selection

Model selected for this project is the Linear regression, it have the smaller RMSE value of 0.065, and because we have available around 400 observations, it is considered important to use other variables like the ones used in the model to make predictions that can allow us to make more assertive decision for marketing investments and other investments that are needed in the Bakery right now, The business at this moment is in the Existence Stage [5], this stage corresponds to the stage where we are looking for customers and developing techniques to respond to these costumers, taking in consideration other variables like website visitors behaviors and social media post's reactions, new promotions and products will a more realistic approach.

Conclusion

Based on the analysis presented before, there are a lot of new decisions to make for marketing strategies based on the graph analysis made, an initial marketing planned activities were detailed before. For the prediction model selected, Linear Regression, it is recommend its deployment by the business, this predictions can show a path for future decisions about investments. This project was made for my own business, I am already using the model for my some of my decisions, based on this, some investments has been planned for the coming months. KNN algorithm was discarded based on value of RMSE. It is important to highlight that these 3 models are different one from the other, one used all numerical independent variables to predict a target variable, the second is an autoregressive model and the third used numerical and non-numerical independent variables to predict the target. For further analysis, the value of the correlation in between Page_views and Site_session and Unique_visitor must be taken in consideration. For the close future, a second part of this project will include social media posts analysis, including Facebook and Instagram, this will complement marketing strategies.

References

[1] Wikipedia contributors. (2021, July 14). Machine learning. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:48, July 25, 2021, from https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=1033602835

[2] Cursera. ARIMA compared to linear regression. Retrieved July 24, 2021, from: https://www.coursera.org/lecture/introduction-trading-machine-learning-gcp/arima-compared-to-linear-regression-ZxJ11

[3] Wikipedia contributors. (2021, July 9). Linear regression. In *Wikipedia, The Free Encyclopedia*. Retrieved 04:33, July 19, 2021, from https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=1032820492

[4] Chauhan, N. Stock Market Forecasting Using Time Series Analysis kdNuggets. Retrieved July 22, 2021, From: https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis.html

[5] Churchill, N. and Lewis, V. (1983). The Five Stages of Small Business Grow. Retrieved on July 25, 2021, from: The Five Stages of Small Business Growth (hbr.org)

Haripriya. R(2019). Sales Predictions using Python for Machine Learninf. Retrieved, July19 , 2021 from: https://hpriya206.medium.com/sales-prediction-using-python-for-machine-learning-6a76e4d63e71

Ahmed, Yacoub.(2019). Predicting stock prices using deep learning. Retrieved, on June 23, 2021 from: https://towardsdatascience.com/getting-rich-quick-with-machine-learning-and-stock-market-predictions-696802da94fe

Luketa. A. Predicting bakery sales with machine learning in Apache Spark. Retrieved, July 18, 2021, from https://xerini.co.uk/predicting-bakery-sales-with-machine-learning-in-apache-spark/

Jim. How to Interpret R-squared in Regression Analysis. Retrieved , July 18, 2021, from https://statisticsbyjim.com/regression/interpret-r-squared-regression/

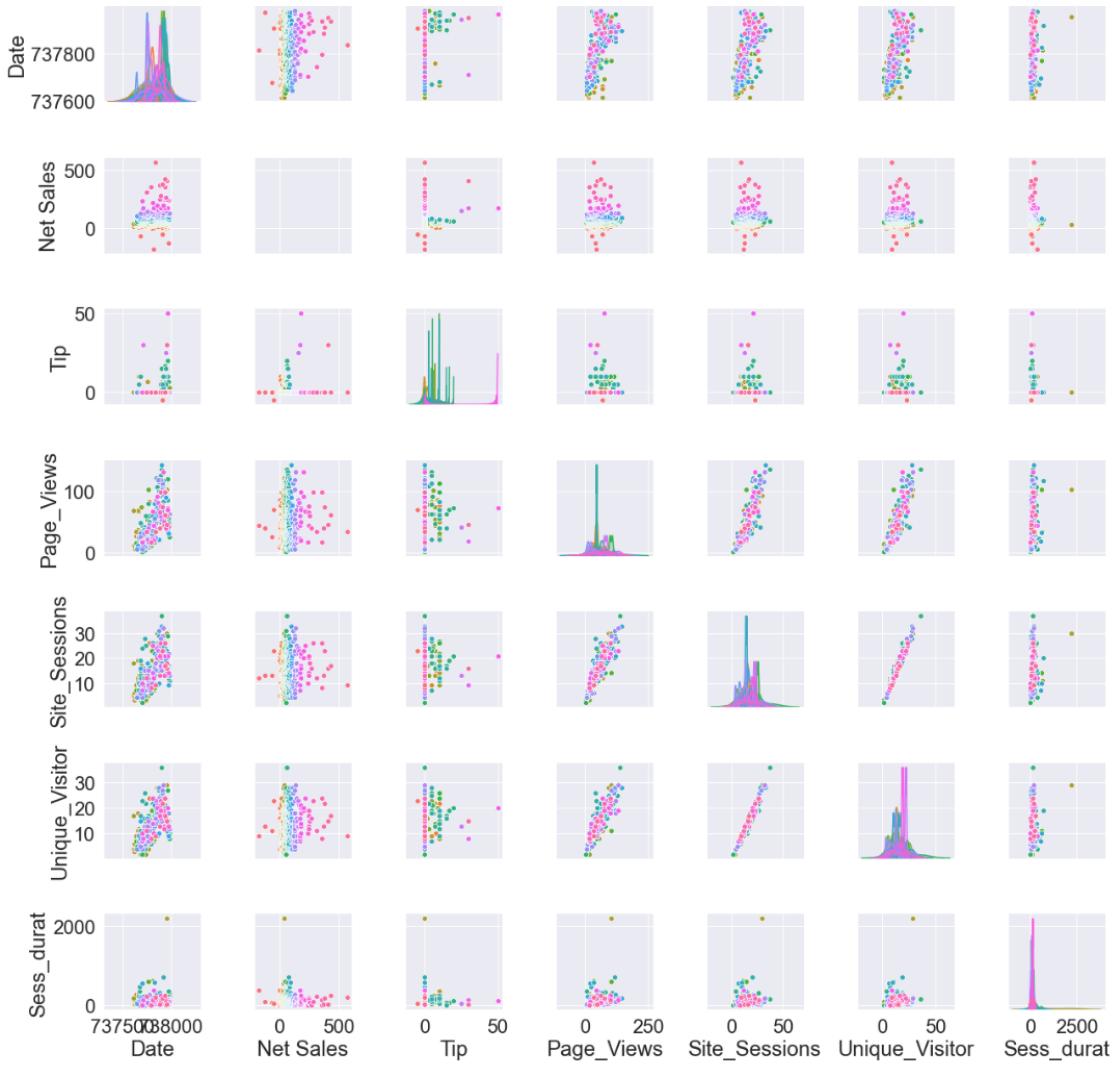
Singh. A. A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code. Retrieve, July 19, 2021, from https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

Kindness, J. (2020). 7 Instagram Metrics You Should Track to Measure Performance. Retrieve, July 15, 2021, from 7 Instagram Metrics You Must Track to Measure Performance - AgencyAnalytics

Watermelon Web Works. LLC. Retrieved, July 10, 2021 from: https://www.watermelonwebworks.com/google-analytics-users-vs-sessions-vs-pageviews/

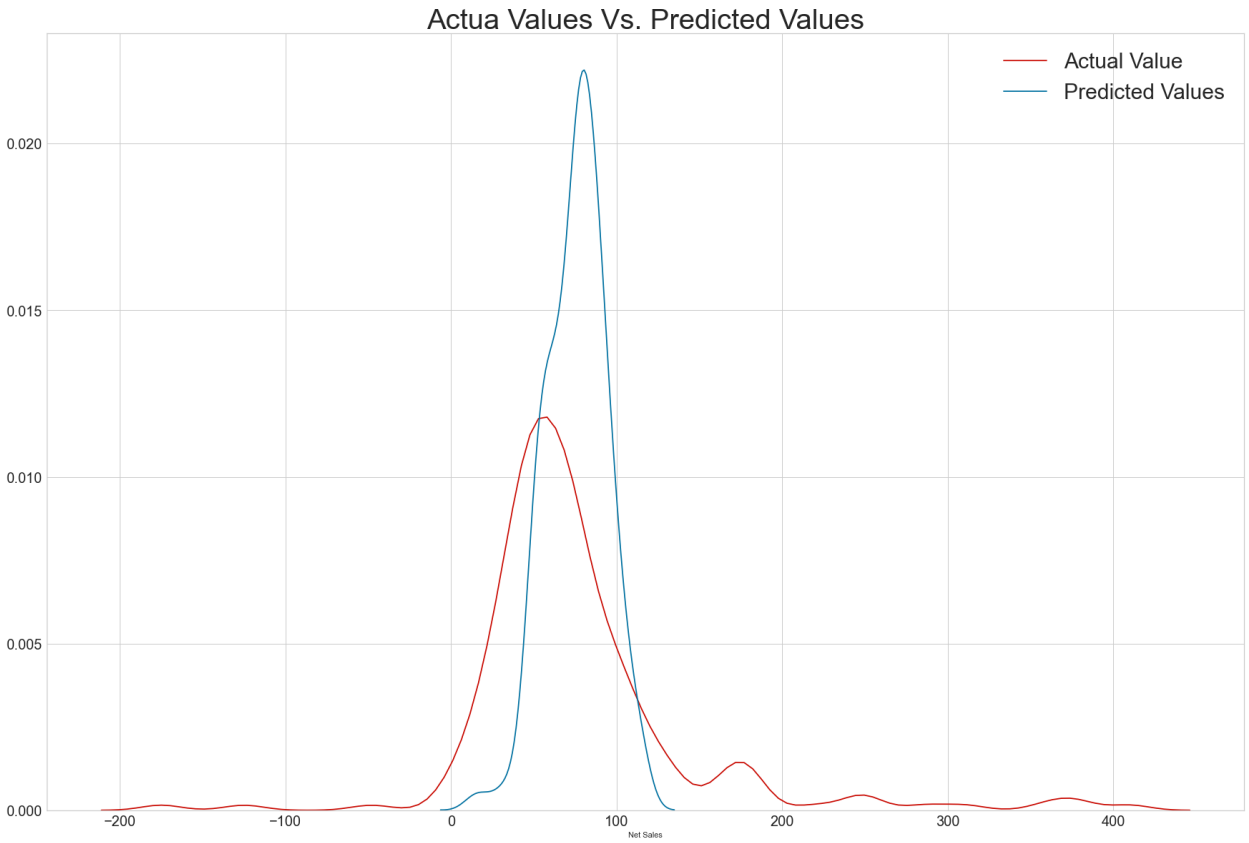
Appendix 1

Scatter Plot Map

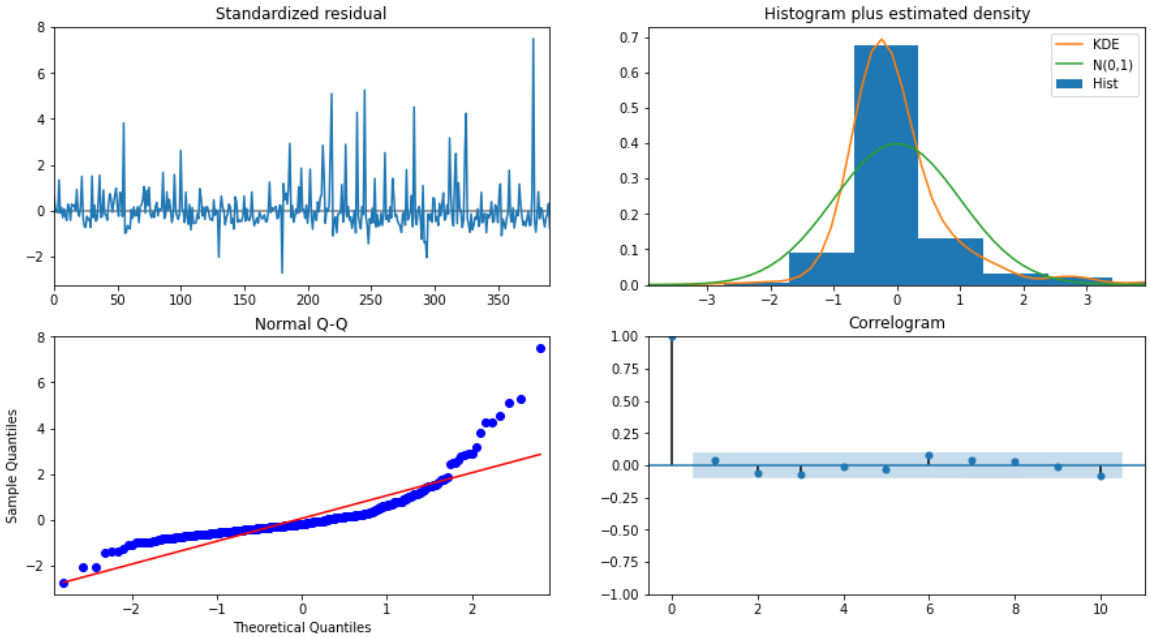


Appendix 2

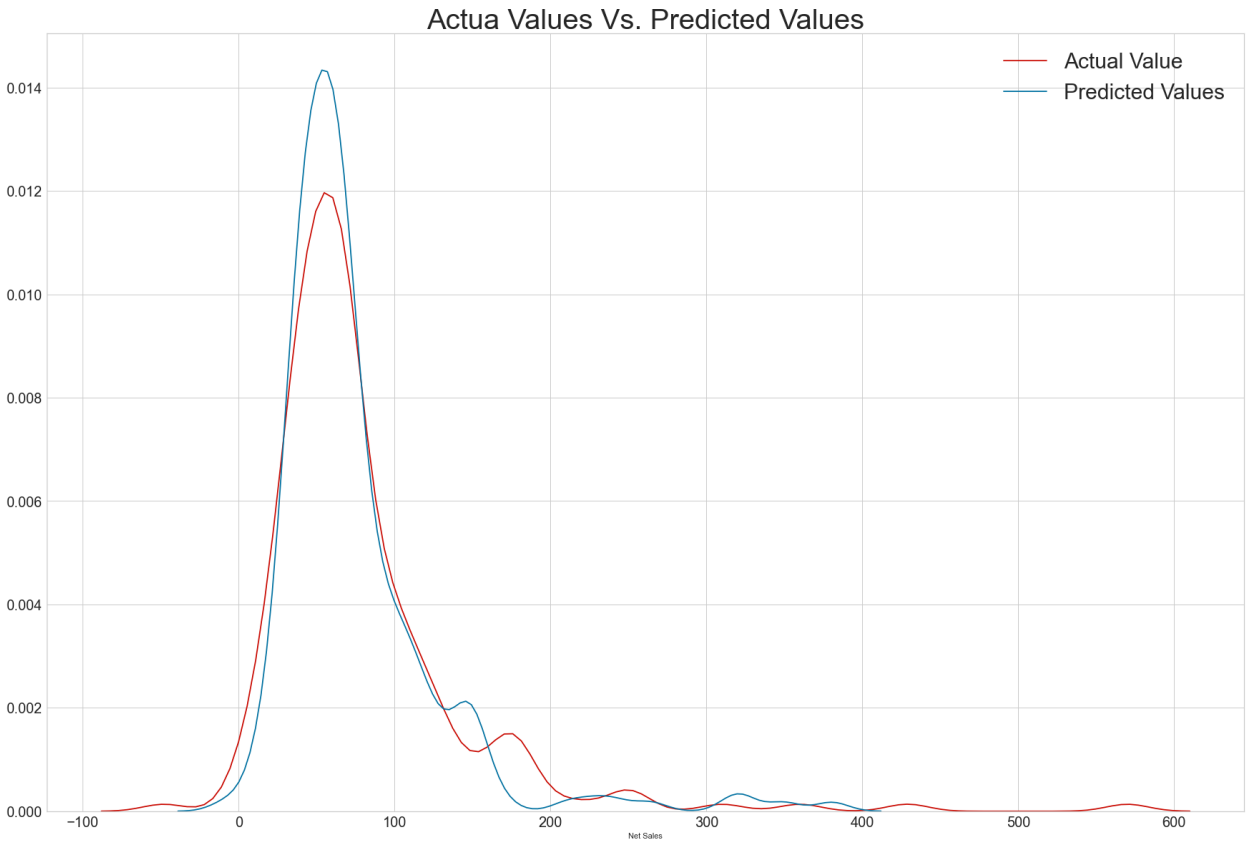
Linear Regression



Appendix 3


Standardized residual
Histogram plus estimated density
KDE
N(0,1)
Hist
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
Correlogram

Appendix 4


Actua Values Vs. Predicted Values
Actual Value
Predicted Values
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−100
0
100
200
300
400
500
600
Net Sales

QUESTIONS

1. **Is there any other model or model optimization technique that can be used to predict sales?**
2. **Why does time of the day is not used as a predictor?**
3. **Is it possible to include in the analysis dates of promotion launches?**
4. **Is there any way to include social media comments or reactions to the predictions?**
5. **How does small data points affect the prediction models?**
6. **Is there any chance to improve performance? How?**
7. **Why the R2 value is so low?**
8. **Can you explain why in the first model the predictions are so different than the actual values?**
9. **What is the process to optimize this model?**
10. **Is there any chance than processing data differently like applying log transformation to the data can improve the results?**