Predictive Cami's Bakery Sales

Gloria P. Moore

DSC680_Applied Data Science

Bellevue University

Bellevue University Data Science Master's Program

Abstract

Predictions is a word that is heard a lot nowadays, predicting sales is a sentence that it is not so new, however is new being commonly used by big number of business executives, how much am I going to sell to be able to develop accurate budget strategies? Where and when am I going to sell? Predicting sales will save a lot of money in time and resources that can be used more effectively if there is an accurate sale prediction available to the decision maker in a company. For this project I am using my own business to predict net sales based on data collected from previous sales and website traffic behavior, using graph analysis to draw a data driven marketing strategy that allow me to increase my sales. The following document will take you through some of my Home-Based bakery data, collected from the invoice system and from the website administrator system. The feature used to predict the Net sales of the bakery were, date, Tip, number of page views, number of site sessions, unique visitor, and duration of the session on my website. Linear regression model was used initially to predict the sales getting a R2 of 0.035, that represent a low value for this model.

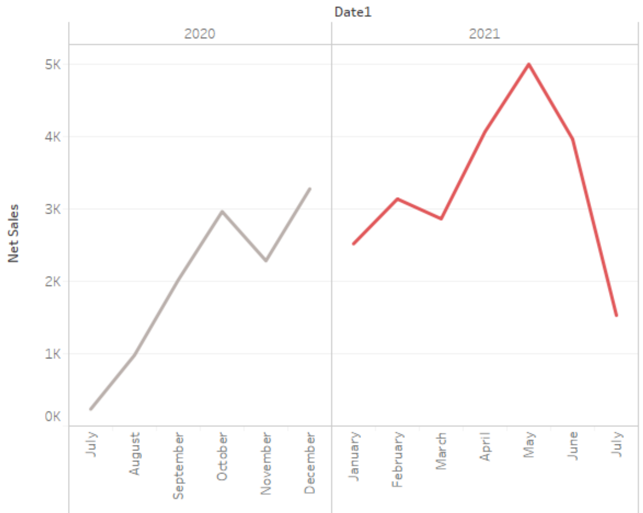
Method

The method used for this project was Linear regression, trying to predict future Net sales for the bakery, the features used for this prediction were:

- Date: The date represent the date were a sale was done
- Tip: Tip left in the sale (US$)
- Number of page views: is the number of times that a single visitor load my page [1]
- Number of site sessions: Number of site sessions, represent a single visit for a specific user, so a page visit can include several sessions [1]
- Unique visitor: each time that a user visits my website for the first time.
- Duration of the session: How long in second a session last for each visitor.

Jupyter notebook was used to collect, clean, organize, process and analyze the data, Pandas library was the most used library for the project, Tableau was used for some of the graphs analysis

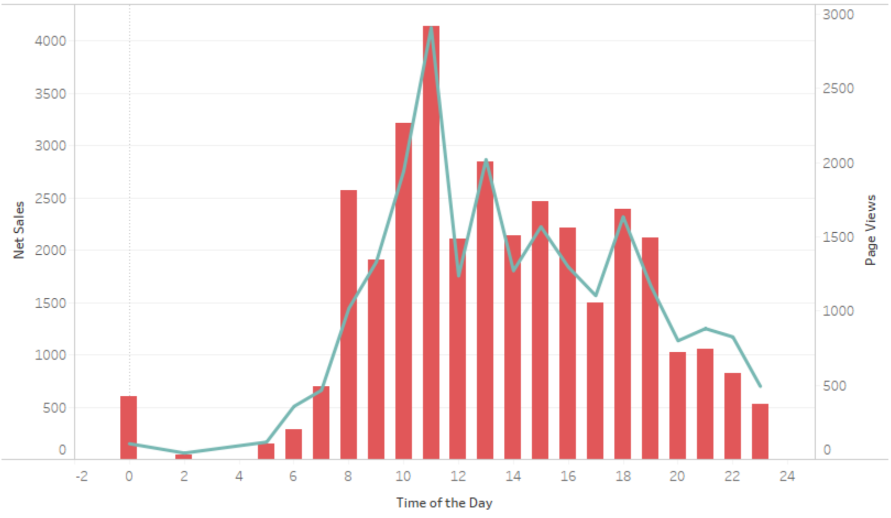
The target variable was defined as Net sales, Gross sales was dropped because it did not have any variation compared with Net Sales, the graph analysis for the target variable include Line graph to compare sales in 2020 and 2021 by month:



Net sales by year

The trend of sum of Net Sales for Date1 Month broken down by Date1 Year. Color shows details about Date1 Year.

Sales was compared with Page Visits by Time of the Day

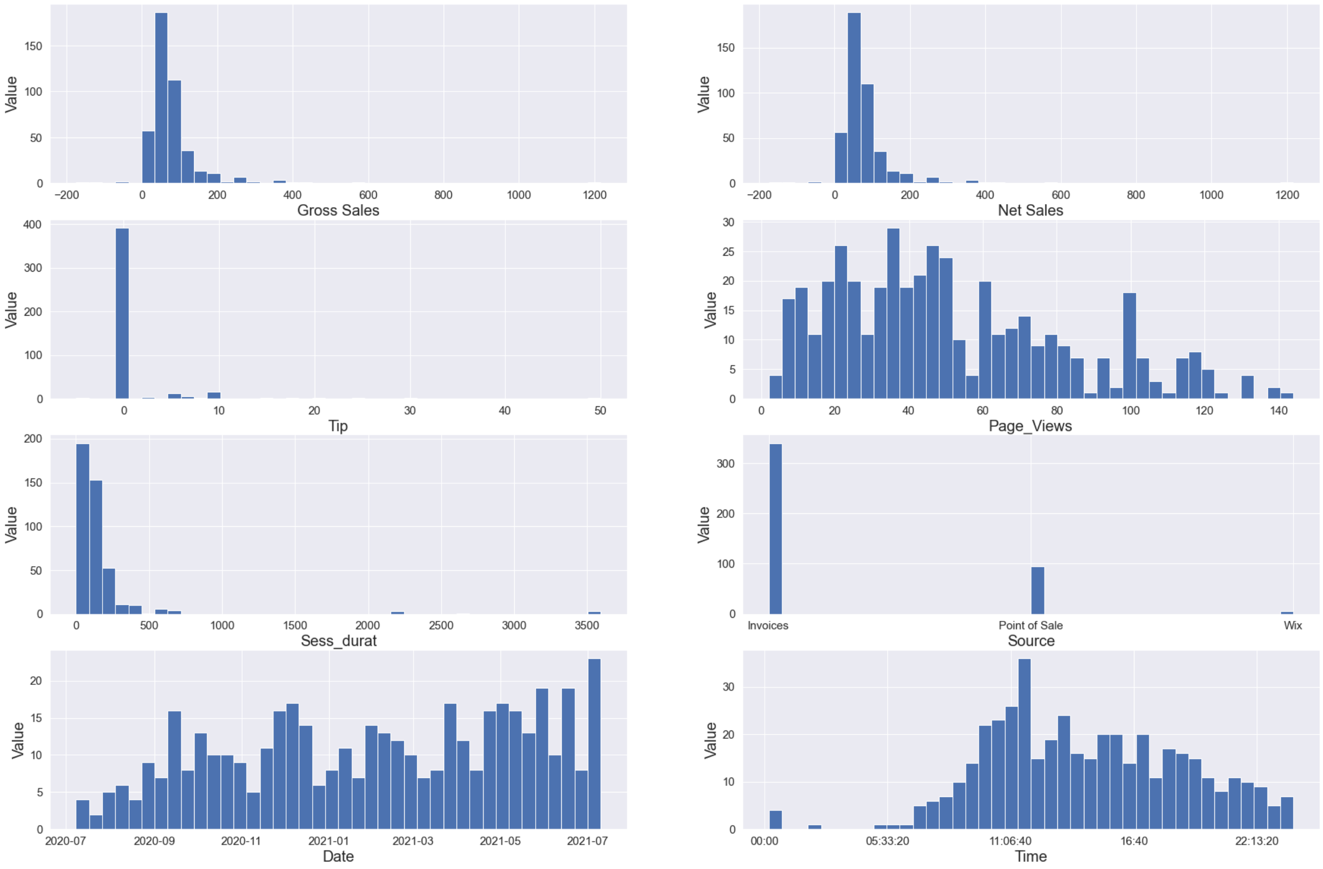


Net Sales and Page Views by Time of the day

The trends of Net Sales and Page Views for Time Hour. Color shows details about Net Sales and Page Views.

Measure Names
🟥 Net Sales
🟩 Page Views

Histograms were created for the features and target variable, from these graphs the Gross Sales column was dropped because it does not present any significant difference from Net sales



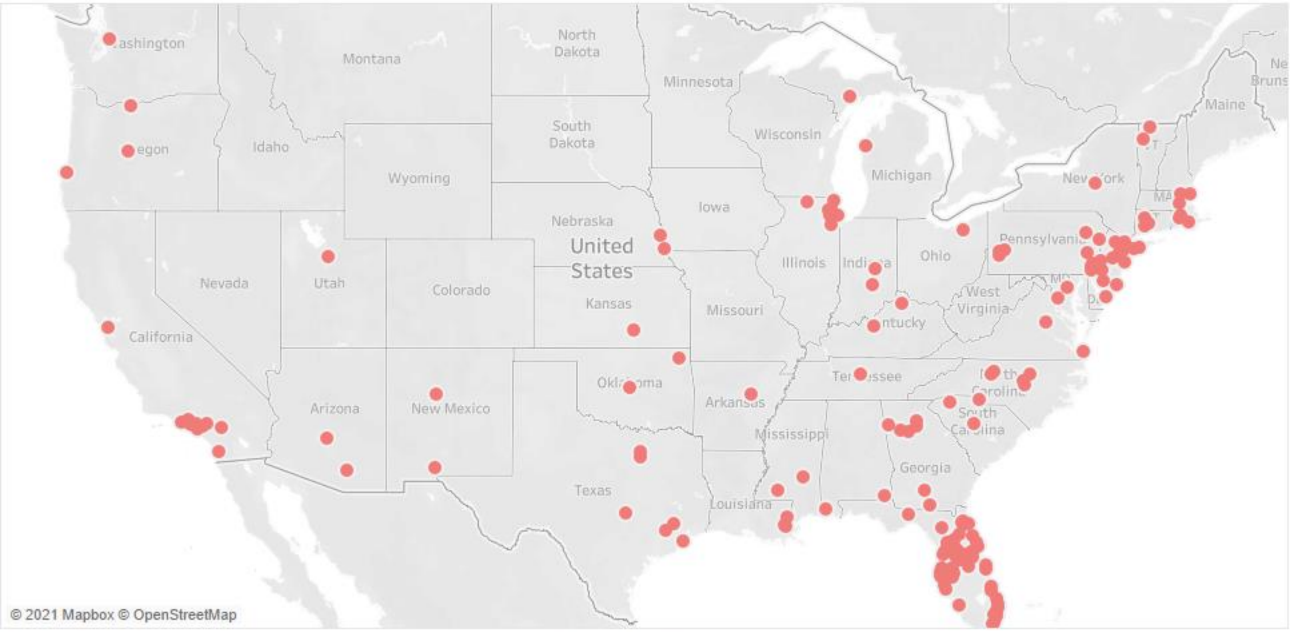
These Histograms were be used to support marketing strategies like:

- Time of the date to post on social media or to promote the website: from 11 am to 4 pm
- 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.

Webiste Traffic by Location US



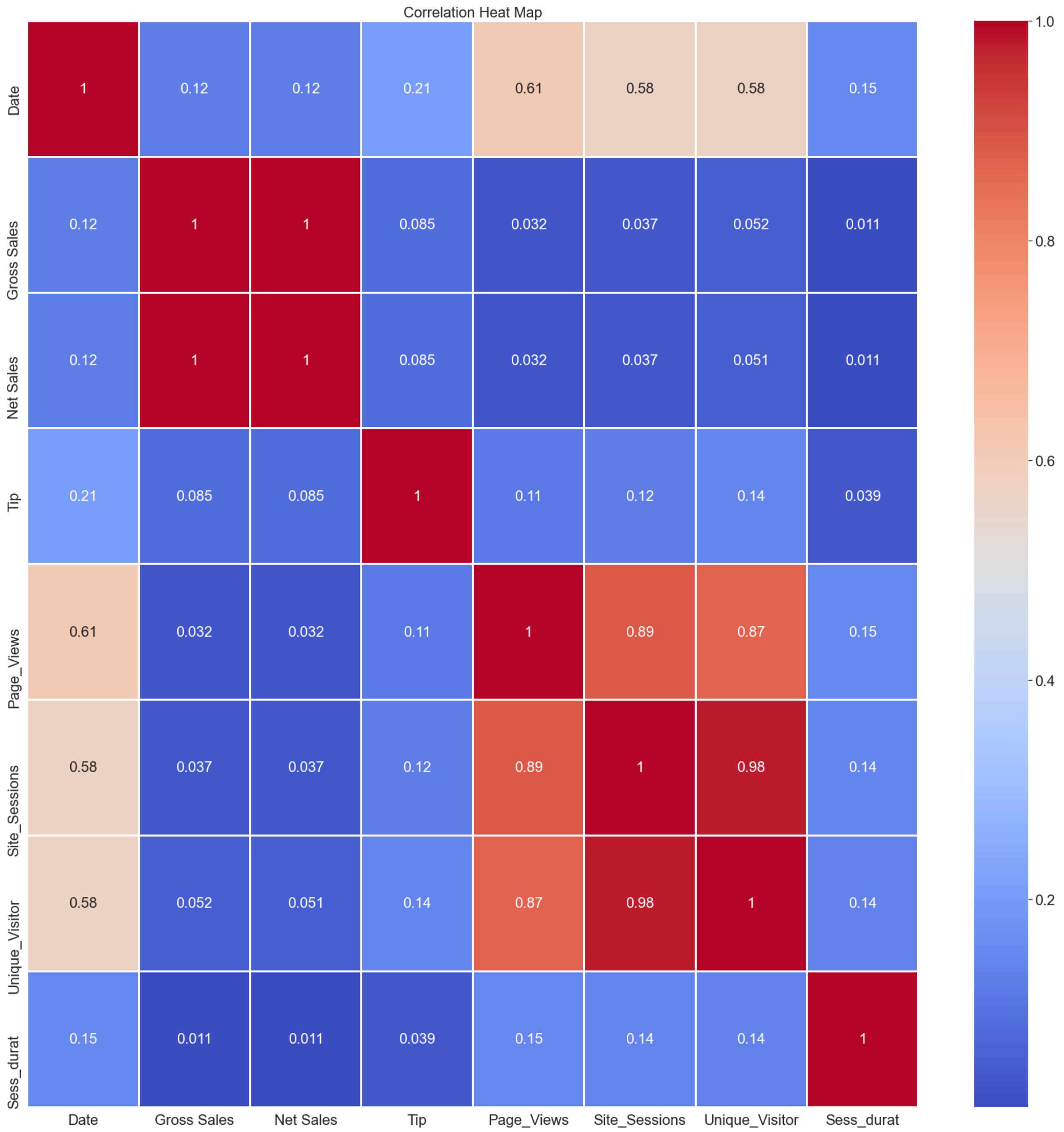
Map based on Longitude (generated) and Latitude (generated). Details are shown for City.

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.

Website visits has been decreasing lately in contrast with the increasing sales (Appendix 1), this phenomenon will be analyzed further in time when we get more data coming from the promotions and marketing done lately.

Modeling

As part of the modeling process, a correlation analysis on the variables was made, and this were the results.

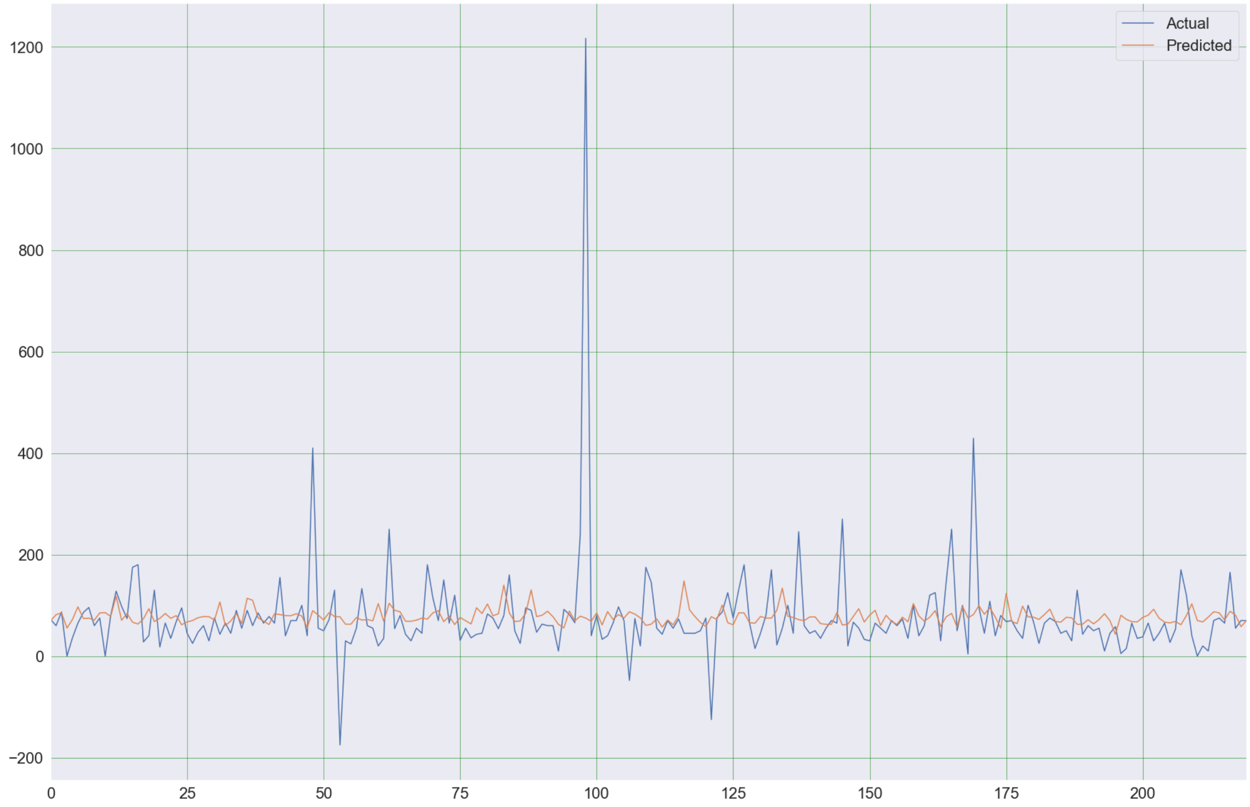


The correlation map does not show any strong correlation values in between variables, it can be concluded that more visitors the website gets, more sales are done, however this correlation values does not show this relationship.

The first attempt of model is a Linear regression with the following results:

```
      Intercept:
 -91924.75741959903
Coefficients:
 [ 0.12470749  2.65828951 -0.2898854  -0.59815036  0.79969485 -0.00391155]
R2:
 0.03539750723692059
```

With an R2 close to 0, this first conclusion can be that this this model does not represent a good model for this data set. Results of Actual vs Predicted Values, code for this graph was adapted from KDNuggets[2]:



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, there are several reprocessing activities over the data like identify possible outlier values of sale in the data set that can be affecting the model. Other processing task will be performed, like location (if enough data is available), and concerning to the model, an optimization will be perform using Principal components and eliminating and adding other predictors, KNN model will be used as well for this prediction[3].

The low value of R2 of 0.035 and the results seen in the graph above shows that this model is not the ideal model to use to predict the Bakery Sales.

References

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

[2] KDNuggets. From Data Pre-processing to Optimizing a Regression Model Performance. Retrieved,July 16, 2021 from: https://www.kdnuggets.com/2019/07/data-pre-processing-optimizing-regression-model-performance.html

[3] 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

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

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/

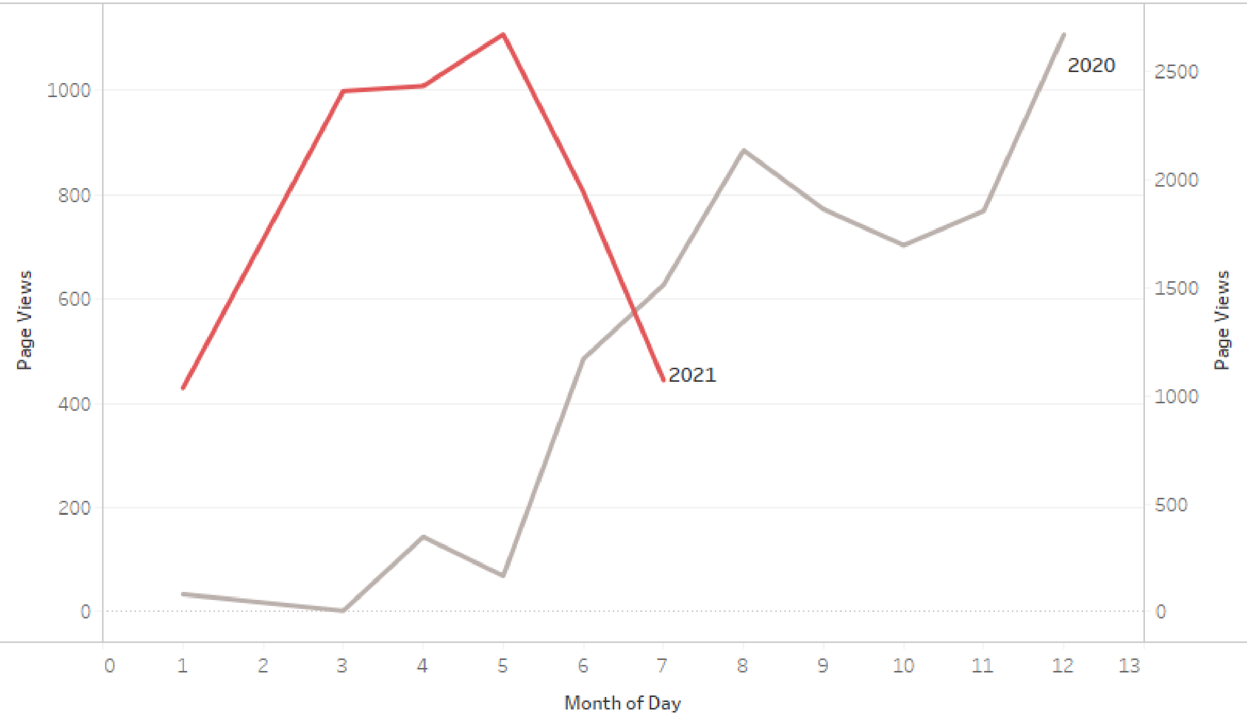
Singh. A. A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code. Retrieve, July 19, 2021, from https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

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

Appendix 1

Graph analysis Page Visits by Year

Page Visits by Year



The trends of Page Views and Page Views for Day Month. Color shows details about Page Views and Page Views.

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?**