Syracuse University

Master of Science in Applied Data Science (ADS)

Portfolio

Sathish Kumar Rajendiran
666555028 (SUID)
srajendi@syr.edu

Overview

- In my current role as a Data Scientist/Sr Manager, I own the Data Strategy & Modernization initiatives at Anaplan and deeply involved in working with data scientists, data analysts, data engineers and business owners across various functional groups. In addition, I am responsible for Data Integration, Data Management, Data Security, Data Quality and Overall Governance of data in compliance with regulatory policies. I have also built the data platform on the Cloud leveraging Big Data technologies by integrating data from transactional systems including as Salesforce, Splunk, Marketo, Gainsight, Anaplan, Greenhouse, Adobe Analytics, Workday, ServiceNow etc. Thus, ensuring a data driven culture for decision makers across various business units.
- I am confident that this master's program in data science would help fine tune my leadership skills, advance my career and abilities to guide, mentor and drive organizations to excel. So, I had signup with this world-renowned accrediting association. Be authentic, think big, do not compromise, and consistently work towards achieving goals are my core principles that has taken this far.

Learning Objective (but not limited to)

- To make data-driven decisions using data capture, management, analysis, and communication
- Understand major practice areas in data science
- Collect, organize, and manage data
- Identify patterns in data using visualization, statistical analysis, and data mining
- Develop actionable insight based on data
- Communicate data analytics and findings to people across a broad range of industries
- Synthesize and understand data science ethics and privacy

Course Structure

- Asynchronous
  - Recorded videos
  - Launch Pad/ Simulation Lab exercises
  - Homework & Diagnostic quizzes
- Synchronous
  - 90 minute - Live sessions
  - Breakout discussions
  - Online Exams

Course Catalog (Total: 36 Credits)

Course Category	Total Credits
Common Core	18
Analytics Applications Core	3
Elective	15

Program	Term	Course 1	Instructor	Start /End Date
Winter 2020	1	MBC 638 - Data Analysis and Decision Making	Prof. Darlene Ryan	01/13/2020 - 03/23/2020
Winter 2020	1	IST 659 - Data Administration Concepts and Database Management	Prof. Gregory Block	01/12/2020 - 03/22/2020
Spring 2020	2	SCM 651 - Business Analytics	Prof. Raghavshyam Ramamurthy	04/06/2020 - 06/15/2020
Spring 2020	2	IST 687 - Introduction to Data Science	Prof. John Santerre	04/07/2020 - 06/16/2020
Summer 2020	3	IST 707 - Data Analytics	Prof.Jermey Bolton	07/05/2020 - 09/12/2020
Summer 2020	3	IST 652 - Scripting for Data Analysis	Prof. D.Landowski	07/02/2020 - 09/10/2020
Fall 2020	4	IST772 - Quantitative Reasoning Data Science	Prof. K Crowston	10/06/2020 - 12/15/2020
Fall 2020	4	IST 664 - Natural Language Processing	Prof. M.Larche	10/05/2020 - 12/14/2020
Winter 2021	5	MAR 653 Marketing Analytics	Prof. R Venkatesan	01/18/2021 - 03/29/2021
Winter 2021	5	IST 719 - Information Visualization	Prof. G.Krudys	01/19/2021 - 03/30/2021
Spring 2021	6	IST 718 - Big Data Analytics	Prof. Jonathan M Fox	04/08/2021 - 06/17/2021
Summer 2021	7	IST 769 - Advanced Database Administration Concepts and Database	Prof. Gregory Block	07/13/2021 - 09/21/2021

» MBC 638 - Data Analysis and Decision Making

- Concepts, principles, and methods to support scientific approach to managerial problem solving and process improvement.
- Basic statistical techniques, their appropriateness to situations and assumptions.

» IST 659 - Data Administration Concepts and Database Management

- Definition, development, and management of databases for information systems.
- Data analysis techniques, data modeling, and schema design.
- Query languages and search specifications.
- Overview of file organization for databases.
- Data administration concepts and skills.

» SCM 651 - Business Analytics

- Business analytics including advanced spreadsheets; relational database and SQL queries.
- statistical analysis in R including multi-linear regression, interactions, tests for regression assumptions, logit, probit.
- neural networks and dashboards.

» IST 687 - Introduction to Data Science

- Introduces information professionals to fundamentals about data and the standards, technologies, and methods for organizing, managing, curating, preserving, and using data.
- Discusses broader issues relating to data management, quality control and publication of data.

» IST 707 - Data Analytics

- General overview in data analytics techniques, familiarity with real-world applications, challenges involved in applications, and future directions of the field.

» IST 652 - Scripting for Data Analysis

- Scripting for the data analysis pipeline. Acquiring, accessing, and transforming data in the forms of structured, semi- structured and unstructured data.

» IST 772 - Quantitative Reasoning Data Science

- Classical statistical procedures used in information transfer research.
- Emphasis on underlying rationale for each procedure and on criteria for selecting procedures in each research situation.

- **IST 664 - Natural Language Processing**
  - Linguistic and computational aspect of natural language processing technologies.
  - Lectures, readings, and projects in the computational techniques required to perform all levels of linguistic processing of text.

- **MAR 653 - Marketing Analytics**
  - Marketing analytics techniques including discriminant analysis, logit, cluster analysis, factor analysis, and conjoint analysis.
  - Marketing decision support models such as new product diffusion, test-market, price, and sales promotion decision models.

- **IST 719 - Information Visualization**
  - A broad introduction to data visualization for information professionals.
  - Develop a portfolio of resources, demonstrations, recipes, and examples of various data visualization techniques.
  -

- **IST 718 - Big Data Analytics**
  - A broad introduction to big data analytical and processing tools for information professionals.
  - Develop a portfolio of theoretical and practical resources for several real-world case studies.

- **IST 769 - Advanced Database Administration Concepts and Database Management**
  - In-depth analysis of relational and non-relational databases and database management system architecture, building complex database objects, database applications using forms and reports, data warehouses, establishing and implementing database security, and tuning databases for optimum performance.

Demonstrate competence and/or mastery of the skills needed

- R
  - Use R to do basic data cleaning and preparation on a wide range of datasets
    - Use functions to summarize and compare fields
    - Find missing values
    - Use subsets or filter data
    - Retype data into correct format
  - Identify stories in datasets through exploration
    - Use R to create appropriate rough plots to identify distributions and relationships in the data
  - Create rich visual artifacts that communicate data stories
    - Identify the optimal type of visualization to minimize viewer cognitive overload and maximize image interpretability
    - Enhance viewer cognition through context cues
    - Use basic design principles to enhance viewer receptivity and convey meaning
    - Develop complex categorical data plot, like alluvial, treemap, and river plots in R

        - Using shape files and geoJSON for map plotting
        - Using Shiny to create interactive dashboards

    - Data Mining & Machine Learning
        - Association Rules, Clustering & Text mining
        - Classification algorithms including, Logistic regression, Support Vector Machines, naïve Bayes, kNN, Random Forest
        - Model evaluation metrics

- Python
    - Exploring and transforming Structured data
        - Arrays, Functions and Categorical summarization
        - Stacking and Unstacking data
        - working with pandas dataframe
    - Semi-structured data
        - NoSQL databases, PyMongo, JSON Encoder and decoder, Processing Twitter, and Facebook APIs
    - Unstructured data
        - Text tokenization, Regular expressions, Finding patterns in Text and Sentiment analysis
    - Network Structures
        - Social network analysis, Geolocations on Maps and analyzing Facebook post comments
    - Visualization
        - Create powerful visualizations using matplotlib, plotly, seaborn, folium, Wordcloud etc.
    - Machine Learning
        - Conventional and Neural network modeling to solve time series, forecasting, regression, image classification and computer vision problems.
        - Designing deep neural network architectures using CNN, RNN models.
    - NLP
        - Linguistic analysis, tokenization, word level semantics, part-of-speech tagging, syntax, semantics, and on up to the discourse level
        - Sentiment Analysis, information extraction (IE), summarization, and machine translation (MT), information retrieval (IR), question answering (QA), and conversational agents

- Relational Databases
    - Programming
        - Creating repeatable scripts, user defined functions, procedures, temporal tables, views etc.
    - Transactions & Concurrency
        - Ensure database transactions are ACID Compliant
    - Performance
        - Analyzing performance of SQL Queries against Big O notation and other performance metrics (index seek vs index scan vs table scans)

        - Working with Clustered, Non-Clustered, Column-store indexes and Indexed views
    - Security
        - Working with permissions and securable
    - Combining relational and NoSQL concepts in SQL Server
        - Evaluating any given database system's architecture fits within the CAP context

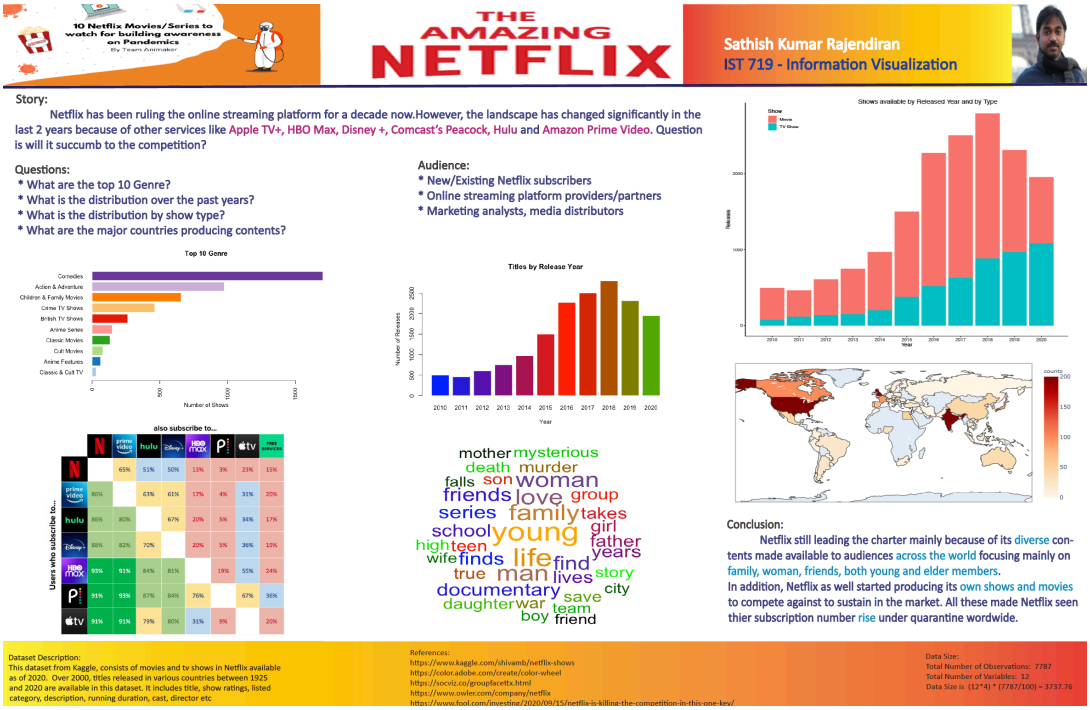
- Big Data
    - Hadoop, HDFS, MapReduce, Hbase, Impala, Pig, Hive, Hcatalog and YARN

- NoSQL (Document, Key-value, column-family, and streaming)
    - MongoDB, Redis, Cassandra, Kafka & KSQL

- Adobe Illustrator
    - Poster Creation

Deliverables/Projects Presented

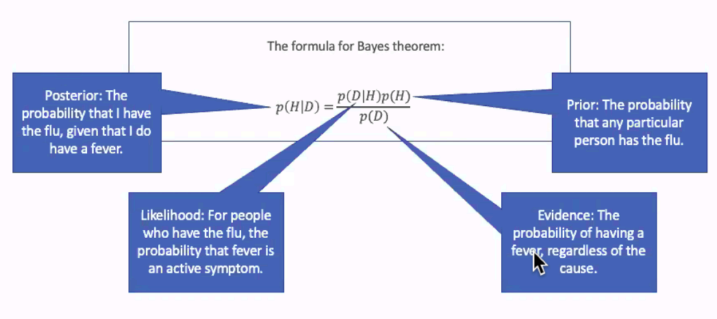
1. **The Amazing Netflix**
    - Description
        - Final Poster presentation – prepare and present an information visualization poster from the dataset chosen.
    - Technology/Tools used:
        - R, Adobe Illustrator
    - Learnings
        - Visually describing a dataset
        - Modifying plots with Illustrator
        - Visual Thinking
            - understanding the difference that makes a difference
        - Visual encoding
        - Visual hierarchy
            - Layouts, Grids & Composition
            - Rule of Thirds & Golden Ratio
        - Grammar of Graphics
        - Poster Critique
    - Source Code
        - https://github.com/sathish-rajendiran/ist719/tree/main/Final_Poster
    - Outcome



2. US Vaccines Data

- Description
  - Using US Vaccines and Schools data – perform
    - exploratory analysis
    - Statistical analysis
      - Null Hypothesis Significance test
      - Confidence Interval
      - Bayesian distribution high density interval
    - Predictive Analysis
      - Build general linear model
      - Multiple linear regression model
      - Bayesian Approach to Multiple Regression Analysis
      - Finally, Comparison to lm (): NHST vs Alternative Hypothesis
- Technology
  - R
- Learnings
  - Explore the characteristics of population using Statistical inferences
  - Population & Sample
    - Sampling distribution
    - Central limit theorem
    - Law of large numbers
    - area under the curve & central region
  - Reasoning with statistical inferences
    - Confidence interval vs Highest density interval
    - Null Hypothesis Significance test (NHST)

- Bayes theorem
    - Likelihood – p(D|H)
    - Prior probability – p(H)
    - Evidence – p(D)
    - Posterior probability – p(H|D)



Bayesian: model the most likely position of the parameter by generating a posterior distribution

Confidence interval: build an interval around a point estimate whose width reflects the uncertainty of that estimate

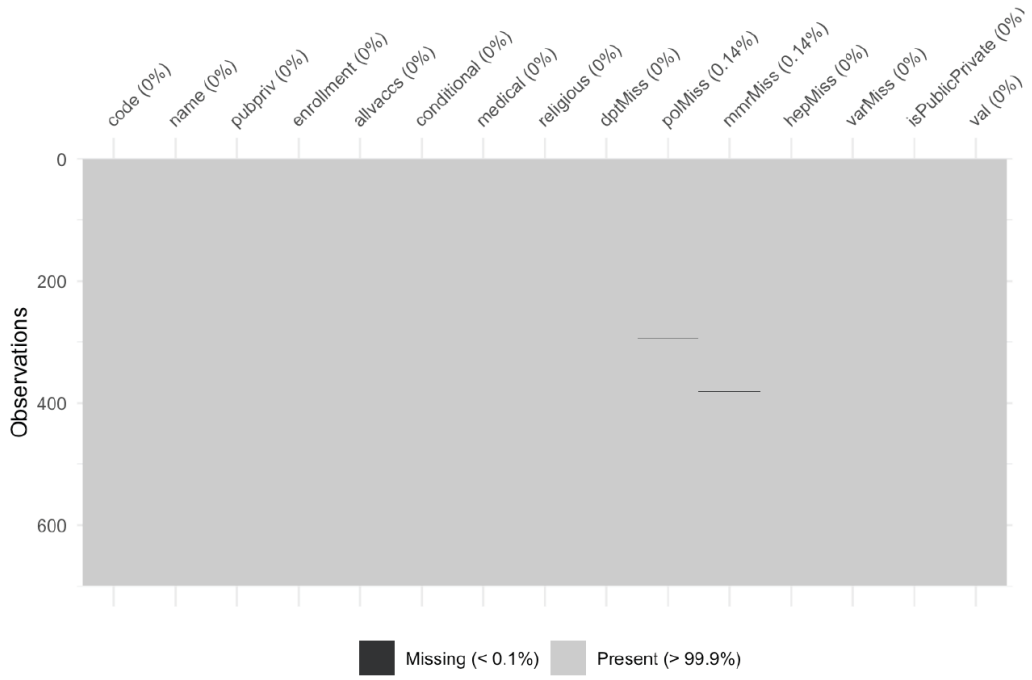
Null Hypothesis Test: make a go/no-go decision by positioning a point estimate on a statistical model of the null hypothesis

- R Squared vs Adjusted R Squared values
- Multicollinearity
- Degrees of freedom
- t-test & F-test
- Time Series components (trend, seasonality, cyclicity, noise and decomposition)

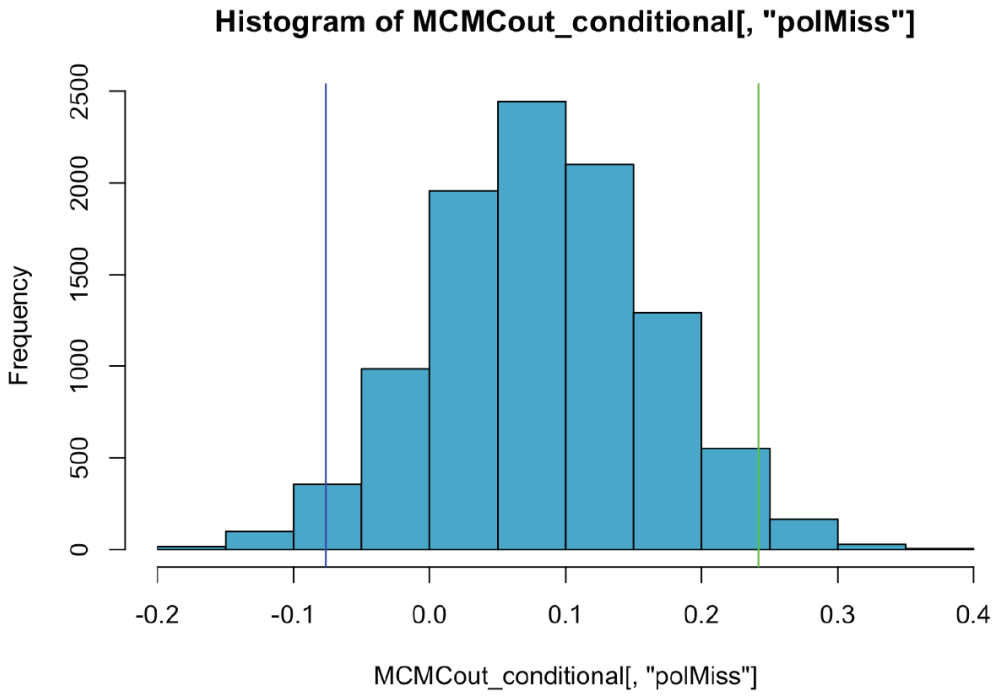
› Source Code
    - https://github.com/sathish-rajendiran/ist772/tree/main/Final_Project

› Outcomes

```
# missing data visualization
vis_miss(reportSampleDF)
```



```
# histogram of 95% HDI mean differences on polMiss | overlaps with 0
hist(MCMCout_conditional[,"polMiss"],col="#4DAFD4")
abline(v=quantile(MCMCout_conditional[,"polMiss"],c(0.025)),col="blue")
abline(v=quantile(MCMCout_conditional[,"polMiss"],c(0.975)),col="green")
```



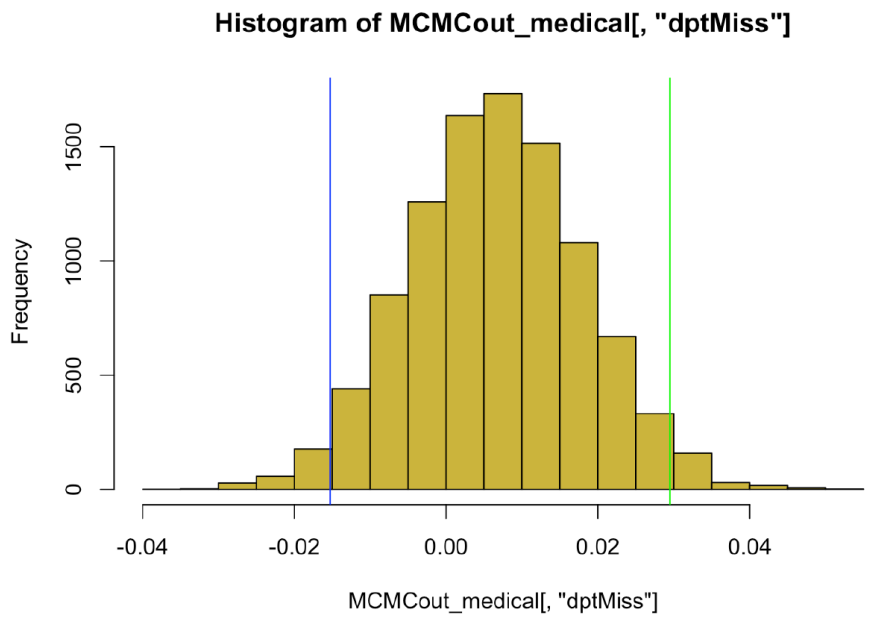

```
medical.lm <- lm(medical ~ dptMiss + polMiss + mmrMiss + varMiss ,data=reportSampleDF)
summary(medical.lm)
```

```
##
## Call:
## lm(formula = medical ~ dptMiss + polMiss + mmrMiss + varMiss,
##     data = reportSampleDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0651 -0.1806 -0.1368 -0.1202 13.9291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.124270   0.043760   2.840  0.00465 **
## dptMiss      0.006927   0.011813   0.586  0.55781
## polMiss     -0.014374   0.012886  -1.115  0.26504
## mmrMiss      0.001800   0.007938   0.227  0.82069
## varMiss      0.017546   0.007232   2.426  0.01552 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8974 on 693 degrees of freedom
## Multiple R-squared:  0.01608,    Adjusted R-squared:  0.0104
## F-statistic: 2.831 on 4 and 693 DF,  p-value: 0.02392
```

```
options(scipen=999)  # turn-off scientific notation like 1e+48
EnsurePackage("BayesFactor")

MCMCout_medical <- lmBF(medical ~ dptMiss + polMiss + mmrMiss + varMiss ,data=reportSampleDF, posterior=TRUE, ite
rations=10000)
summary(MCMCout_medical)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##              Mean       SD   Naive SE Time-series SE
## mu       0.176513 0.034189 0.00034189     0.00036380
## dptMiss  0.006635 0.011378 0.00011378     0.00011378
## polMiss -0.013715 0.012598 0.00012598     0.00012598
## mmrMiss  0.001726 0.007681 0.00007681     0.00007681
## varMiss  0.016715 0.007136 0.00007136     0.00007104
## sig2     0.803768 0.043340 0.00043340     0.00043612
## g        0.049418 0.069260 0.00069260     0.00069260
##
## 2. Quantiles for each variable:
##
##              2.5%       25%       50%       75%   97.5%
## mu       0.109806  0.153174  0.176351  0.199960 0.24327
## dptMiss -0.015334 -0.001110  0.006665  0.014246 0.02944
## polMiss -0.038650 -0.022041 -0.013680 -0.005383 0.01132
## mmrMiss -0.013355 -0.003397  0.001663  0.006854 0.01689
## varMiss  0.002626  0.011880  0.016714  0.021498 0.03072
## sig2     0.723826  0.773756  0.802510  0.832234 0.89318
## g        0.011025  0.021784  0.033495  0.055239 0.17953
```

```
# histogram of 95% HDI mean differences on dptMiss | overlaps with 0
hist(MCMCout_medical[,"dptMiss"],col="#CBB43D")
abline(v=quantile(MCMCout_medical[,"dptMiss"],c(0.025)),col="blue")
abline(v=quantile(MCMCout_medical[,"dptMiss"],c(0.975)),col="green")
```

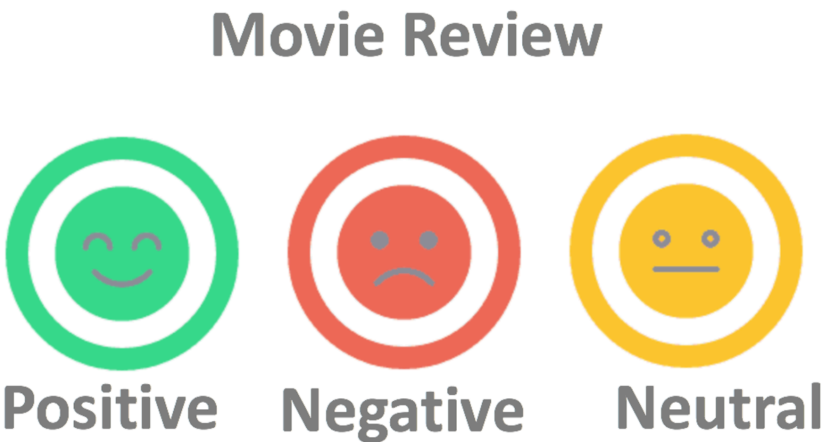


3. Sentiment Analysis on Movie Reviews

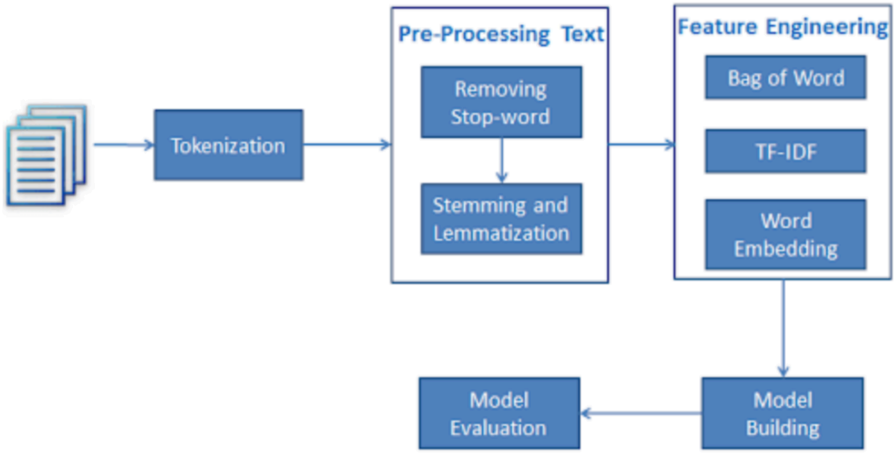
- Description

  In this project, our objective is to classify the movie reviews as either positive, negative, or neutral from the Kaggle movie reviews dataset using NLTK libraries in Python. It involves, Data collection, Pre-processing, Feature engineering, and experimentation using NLTK classifier algorithms. Our expectation is to try variety of models and preprocessing techniques and choose an optimal combination with accuracy expected in the ranges of 60 to 80%.

- Technology
  - NLTK Python library
- Learnings
  - ZIPf's Law
    - The frequency of any word is inversely proportional to its rank in frequency table *f r=k (for constant k)*
  - Language models
  - Regular expressions
  - Morphology
    - Lexical, Syntactic, Semantic, Discourse & Pragmatic
    - Inflection, Derivation, Stemming, Lemmatization
  - Tokenization
  - Parts of Speech
  - Classification techniques
    - decision tree, rule-based, memory based, instance based, support vector machines, naïve bayes, neural networks and genetic algorithms
  - Classifier evaluation methods
    - confusion matrix
    - precision, recall, F-measures
  - Issues with imbalanced classes
  - Sentiment Analysis
  - Information Retrieval
- Source Code
  - https://github.com/sathish-rajendiran/ist664/tree/main/Final_Project
- Highlights



NLTK Text Analytics



```
#plot the summary
plt.rcParams['figure.figsize'] = (13, 7)
sns.countplot(train["Sentiment"], palette='rocket')
plt.title('Sentiment Score in Training Dataset', fontsize = 20)
```

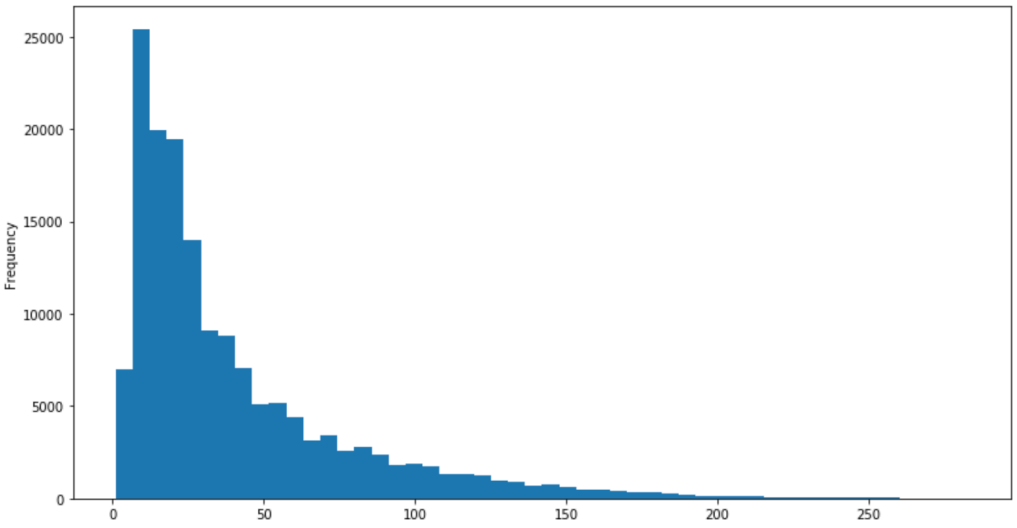
```
Number of Postive Sentiment:  42133
Number of Neutral Sentiment:  79582
Number of Negative Sentiment:  34345

Percentage of positive Sentiment 27.0%
Percentage of neutral Sentiment 50.99%
Percentage of negative Sentiment 22.01%
```

```
Text(0.5, 1.0, 'Sentiment Score in Training Dataset')
```

Analyze the spread of review by Number of characters

```
train['length'].plot(bins=50, kind='hist')
```

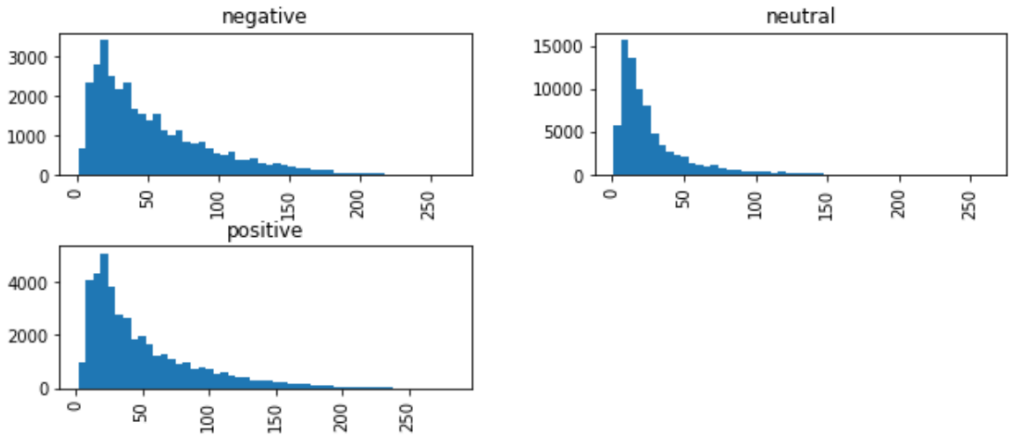
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1ee6da50>
```



Analyze the spread of reviews by Score/Sentiment

```
train.hist(column='length', by='Score', bins=50,figsize=(10,4))
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x1a224f9850>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x1a1ee097d0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x1a1e85cfd0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x1a1e899810>]],
      dtype=object)
```




```
# move common words with frequency into temp df
temp = pd.DataFrame(clean_reviewsFD.most_common(30))
temp.columns = ['Common_words','count']
temp.style.background_gradient(cmap='Blues')
```

```
16387
156060
```

:

	Common_words	count
0	one	3609
1	like	3071
2	story	2520
3	good	2043
4	characters	1882
5	much	1862
6	time	1747
7	comedy	1721
8	even	1597
9	little	1573
10	funny	1522
11	way	1511
12	life	1484

```
# execute the model
model.fit(X_train, y_train, batch_size=128, epochs=7, verbose=1)
```

```
Epoch 1/7
976/976 [==============================] - 105s 108ms/step - loss: 0.9936 - accuracy: 0.6015
Epoch 2/7
976/976 [==============================] - 104s 106ms/step - loss: 0.8064 - accuracy: 0.6706
Epoch 3/7
976/976 [==============================] - 104s 107ms/step - loss: 0.7416 - accuracy: 0.6930
Epoch 4/7
976/976 [==============================] - 108s 111ms/step - loss: 0.6916 - accuracy: 0.7108
Epoch 5/7
976/976 [==============================] - 117s 120ms/step - loss: 0.6526 - accuracy: 0.7254
Epoch 6/7
976/976 [==============================] - 116s 119ms/step - loss: 0.6211 - accuracy: 0.7361
Epoch 7/7
976/976 [==============================] - 113s 116ms/step - loss: 0.5963 - accuracy: 0.7439

<tensorflow.python.keras.callbacks.History at 0x1a3053fd50>
```

Tree of Most Common Words

Classifier	Precision	Recall	F1-Score	Support	Accuracy
Multinomial Naïve Bayes	0.79	0.64	0.68	31212	0.64
Multinomial Naïve Bayes with Lemmatization	0.65	0.61	0.63	31212	0.61
Multinomial Naïve Bayes with Stemming	0.66	0.61	0.63	31212	0.61
Naïve Bayes with Cross Validation: Fold 10	0.56	0.53	0.53		
Naïve Bayes with Cross Validation: Fold 5	0.56	0.53	0.53		
Random Forest	0.73	0.71	0.72	31212	0.71
Tensorflow					0.74

Clearly, Tensorflow had the best result with almost ~75% accuracy.

	Multinomial Naïve Bayes			
Sentiment	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.44	0.76	0.56	4828
Negative	0.27	0.75	0.39	2487
Neutral	**0.91**	0.61	0.73	23897
Overall Accuracy		**0.64**		

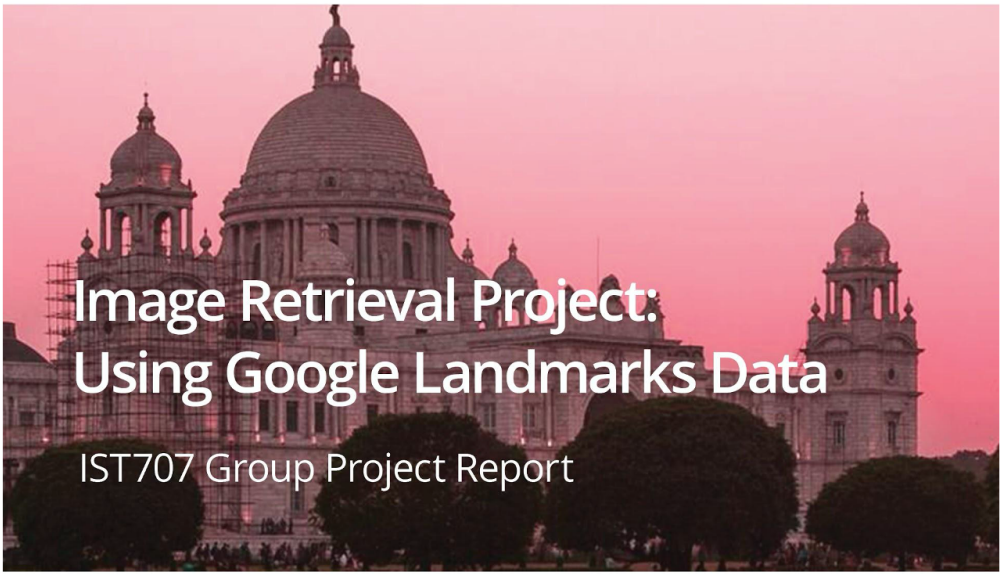
	Random Forest			
Sentiment	**Precision**	**Recall**	**F1-Score**	**Support**
Positive	0.63	0.74	0.68	6995
Negative	0.57	0.73	0.64	5463
Neutral	**0.82**	0.7	0.76	18754
Overall Accuracy		**0.71**		

4. Computer Vision on Google Landmarks Data

- Description

  The project team has chosen the data set that comes from the interesting Kaggle competition “Image retrieval using Google Landmarks data”. The objective was to predict the images correctly by labels, using deep learning techniques (convolutional neural networks) and to compare the results with traditional models, such as Decision tree, k-Nearest Neighbors, Support Vector Machines and Random Forest. In addition, the team investigated the configuration of the architecture of convolutional neural networks (CNNs) and explored hyperparameter-tuning for model-fitting to determine whether CNNs are better than traditional classification algorithms for an image classification task.

- Technology
  - R & Weka
- Learnings
  - Data mining tasks
    - Classification, Clustering and Association rule mining
  - Description vs Prediction
  - Correlation vs Causation
  - Data Transformation
  - Sampling methods
  - Association rule
    - Support, Confidence and Lift
    - Apriori principle
  - Clustering
    - Partitional vs Hierarchical
    - Distance measures
  - Model evaluation
    - Overfitting, Hold-out, cross validation
    - Precision, recall & F-measure
    - Speed, robustness, scalability & Interpretability
  - Bayes theorem
    - Conditional vs joint probabilities
    - Naïve Bayes classifier
  - k-Nearest Neighbor
  - Support vector machines
- Source Code
  - https://github.com/sathish-rajendiran/ist707/tree/main/Final_Project
- Outcomes



Professor

Jeremy Bolton

Project Team

Daphne Chang
Sathish Kumar Rajendiran
Sharat Sripada

IST707 Data Analytics

School of Information Studies
Syracuse University

Example of landmark diversity below:

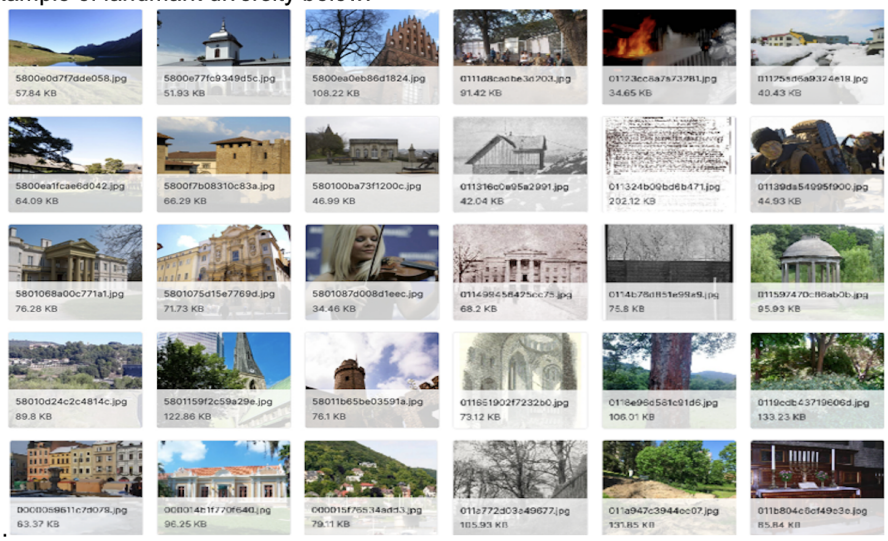


Fig: Examples from the original data set

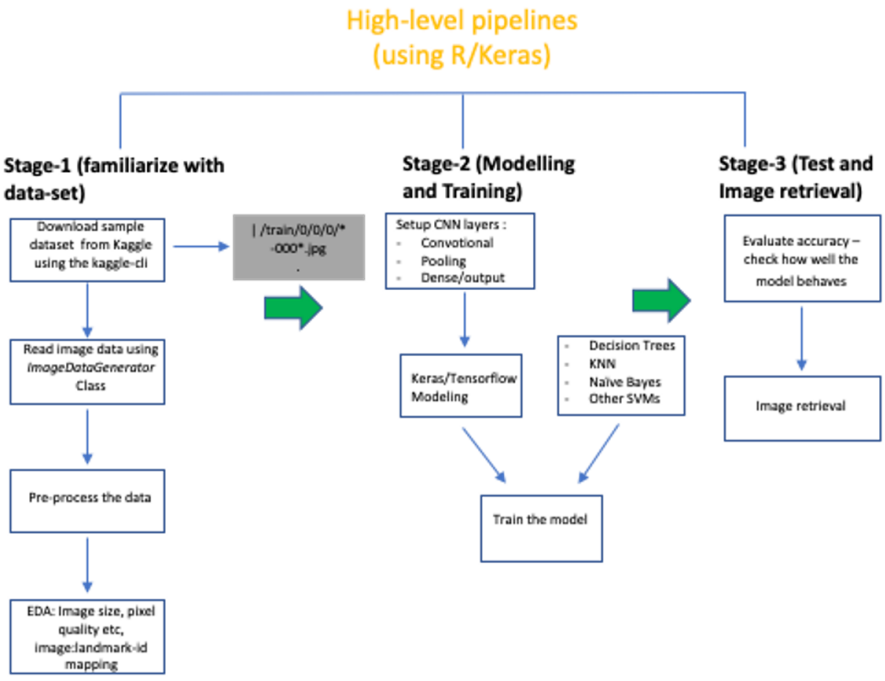


Fig: Work process pipelines

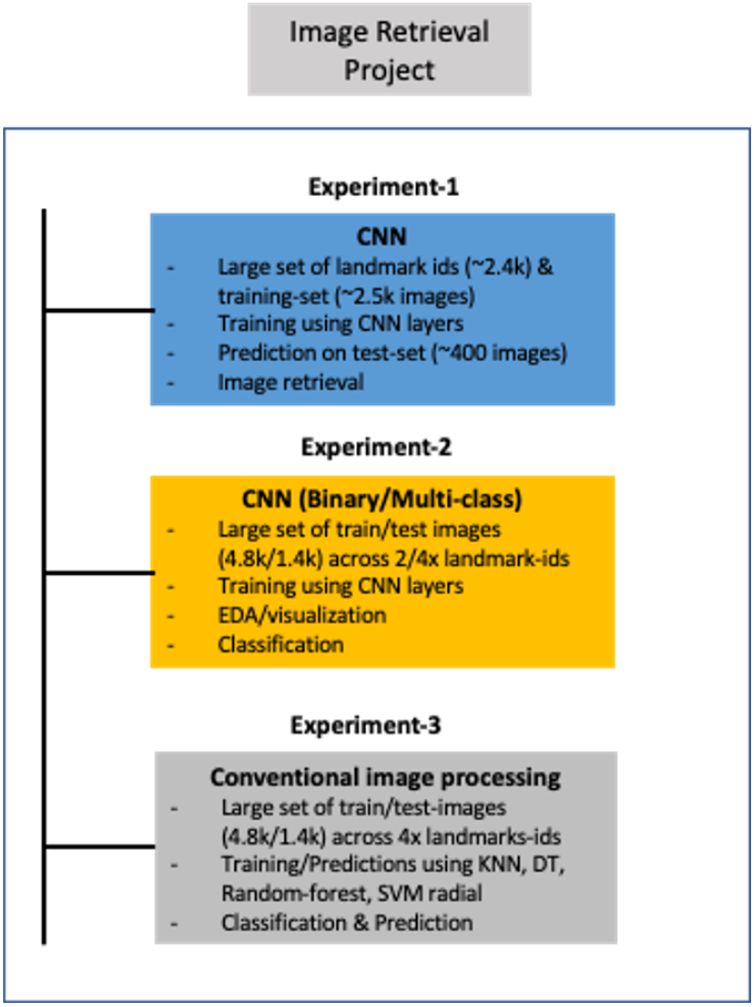


Fig: High-level project execution strategy

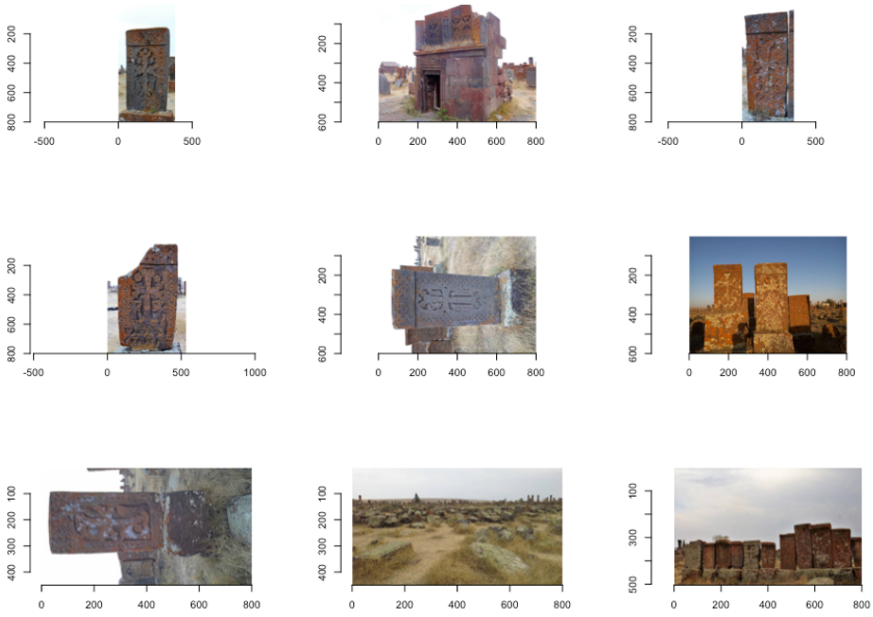
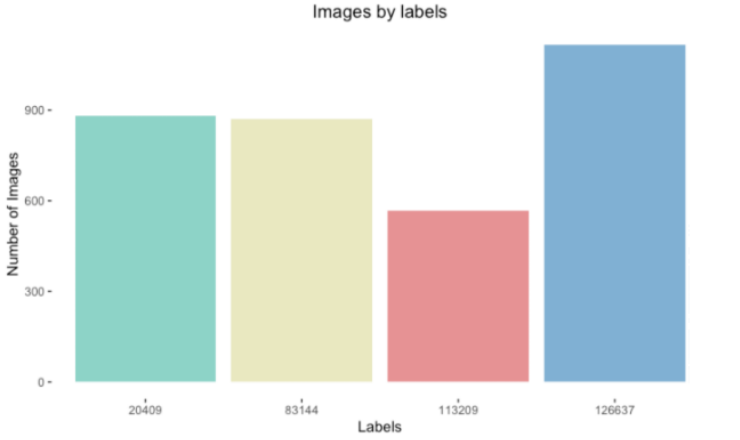


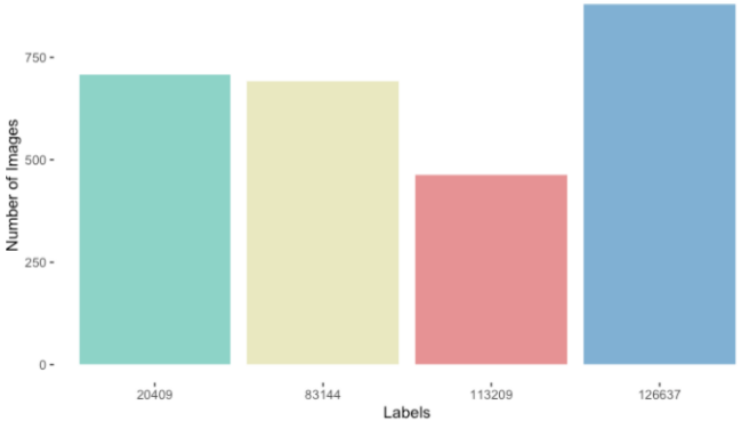
Fig: Sample Images with Landmark ID 20409

Fig: Activation Maps on Intermediate Layers of Binary Classification Model Two

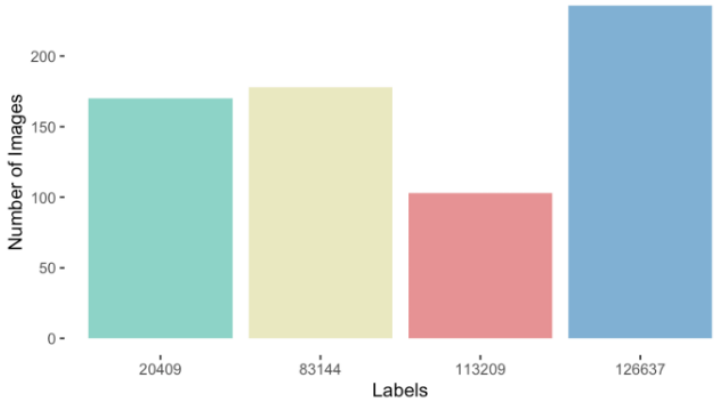
Sample data is further split into train and test data with 80:20 ratio as below,

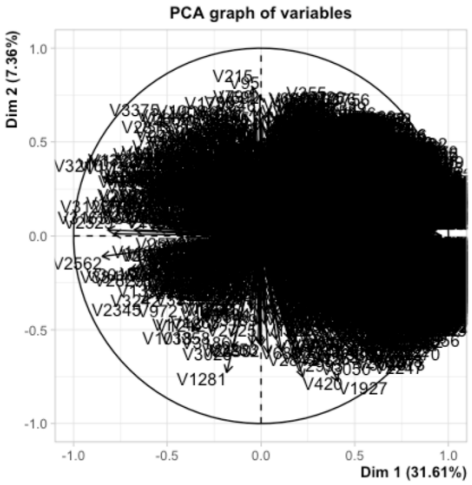


Train data images by labels

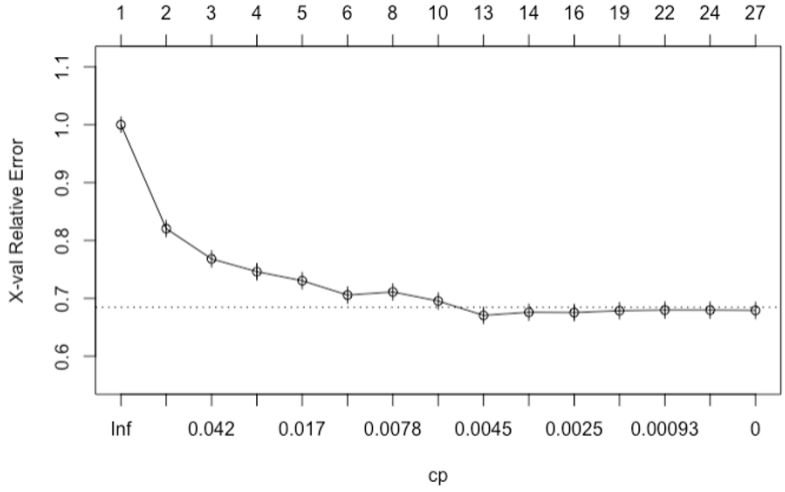


Test data images by labels

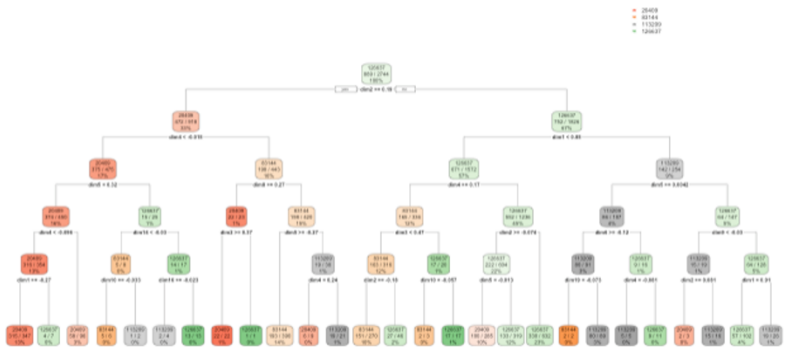





```
#Rename Columns
colnames(imagePCATrainDF) <- c("label","dim1","dim2","dim3","dim4","dim5","dim6","dim7","dim8","dim9","dim10","dim11","dim12","dim13"
                               ,"dim14","dim15","dim16","dim17","dim18","dim19","dim20","dim21","dim22","dim23","dim24","dim25"
                               ,"dim26","dim27","dim28","dim29","dim30")
str(imagePCATrainDF)
dim(imagePCATrainDF)
```



Classification Tree for Images after PCA



Conclusions

The landmark retrieval problem is a paradigm in itself. While there is a steep learning curve to image processing with understanding the architecture of Convolutional Neural Networks (CNN) there are deeper aspects to experimentation at scale. Given the problem complexity at the outset and the resource limitations (lack of compute resources or GPUs), the Team dissected the project into three parts:

- Focusing on exploring CNNs, generators to sift through images at scale and experimentation with convnet parameters to improve prediction
- Deeper understanding of image processing at each layer and other aspects of CNNs while limiting the data-set classifications to binary or multi-class at most
- Exploring traditional classification and prediction algorithms for image processing

Details of datasets, excerpts of code, meaningful outputs and analysis has been presented in an attempt to fully explain the problem at hand and tools that were used to solve it. The findings from the study are:

- For image classification, convolutional neural networks give much higher accuracy in making predictions than traditional classification algorithms, as shown in the table below.
- CNNs can even work well with very small sample sets.
- When dealing with large image data sets, like Google Landmark data, CNNs need much computing power to carry out their "learning".
- Keras library with TensorFlow backend provides an easy-to-use framework that allows fast prototyping.
- Hyperparameter-tuning of CNNs can get very complex and time-consuming.

Accuracy Comparison for All Models

Model	Accuracy (4 labels)	Accuracy (2 labels)
Decision Tree	53.71%	70%
Random Forest	63.54%	83.70%
KNN	56.73% (k=5)	80.34% (k=17)
SVM Radial	61.43%	78.86%
CNN	93.1%	97.7%

5. Traffic Collision Analysis

› Description

As the second largest in the United States, Los Angeles has traffic challenges due to a large and growing population and an increase in the number of cars. A better understanding of the factors that contribute to accidents can help government officials, companies, citizens, and other interested parties to understand how to make the city safer and more drivable.

The goal is to explore the trends and correlations between the data to provide useful information that can help answer our proposed analysis questions:

1. What are the most dangerous intersections?
2. What are the most common collision areas in Los Angeles?
3. What are the best/worst times of the day for accidents? Best/worst month?
4. What is the demographic makeup of victims in collisions?

5. What is the relationship between income and collision victims?
6. Do certain temperatures or weather play a factor?

- Technology
  - Python
- Source Code
  - https://github.com/sathish-rajendiran/ist652/tree/main/Project/Final_Submission
- Outcome

The Los Angeles Traffic Collision Data is publicly available from Kaggle.com is owned by the City of Los Angeles. The contains 481,568 incidents from 2010 to 2019.

Source: https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data

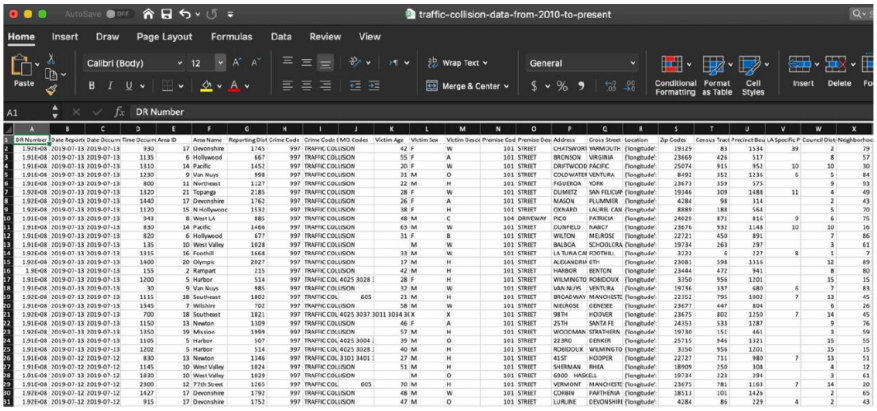
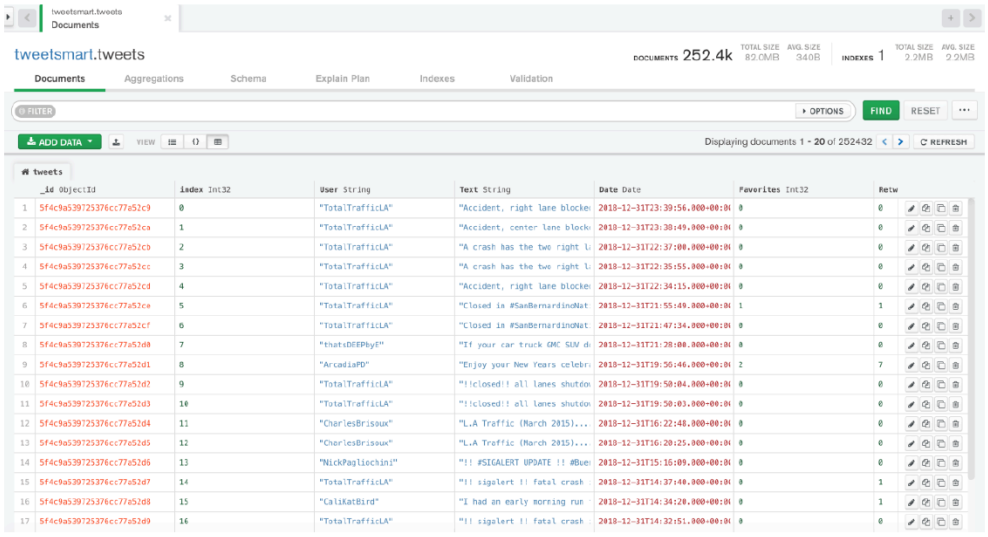


FIGURE 1 - LOS ANGELES TRAFFIC DATA 2010-2019 FROM KAGGLE.COM

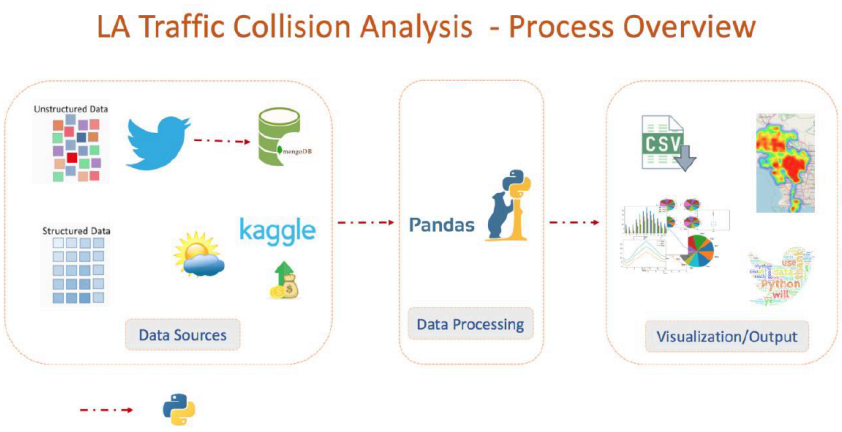
Multiple data sets were incorporated into analysis beyond the original Kaggle dataset in order to get more data that was not included in this original dataset such as weather and income data.




```
{
    "_id": {
        "$oid": "5f4c9a539725376cc77a52c9"
    },
    "index": 0,
    "User": "TotalTrafficLA",
    "Text": "Accident, right lane blocked in #MorenoValley on 60 EB before Heacock St, stopped traffic back to I-215, delay of 24 mins #LAtraffic",
    "Date": {
        "$date": "2018-12-31T23:39:56.000Z"
    },
    "Favorites": 0,
    "Retweets": 0,
    "Mentions": "",
    "Hashtags": "#MorenoValley #LAtraffic",
    "Geolocation": ""
}
```

```
{
    "_id": {
        "$oid": "5f4c9a539725376cc77a52ca"
    },
    "index": 1,
    "User": "TotalTrafficLA",
    "Text": "Accident, center lane blocked in #SanBernardino on I-215 SB before Mill St, stopped traffic back to Hwy 66, delay of 9 mins #LAtraffic",
    "Date": {…},
    "Favorites": 0,
    "Retweets": 0,
    "Mentions": "",
    "Hashtags": "#SanBernardino #LAtraffic",
    "Geolocation": ""
}
```


PROCESS OVERVIEW



The location was analyzed in many ways. Heat maps of the collisions were made to show patterns of where the concentration of collisions in Los Angeles occur.

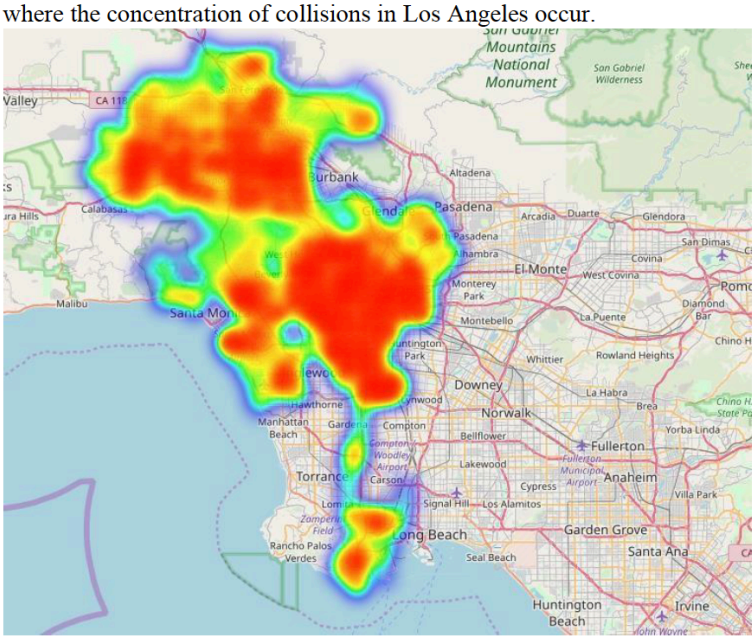
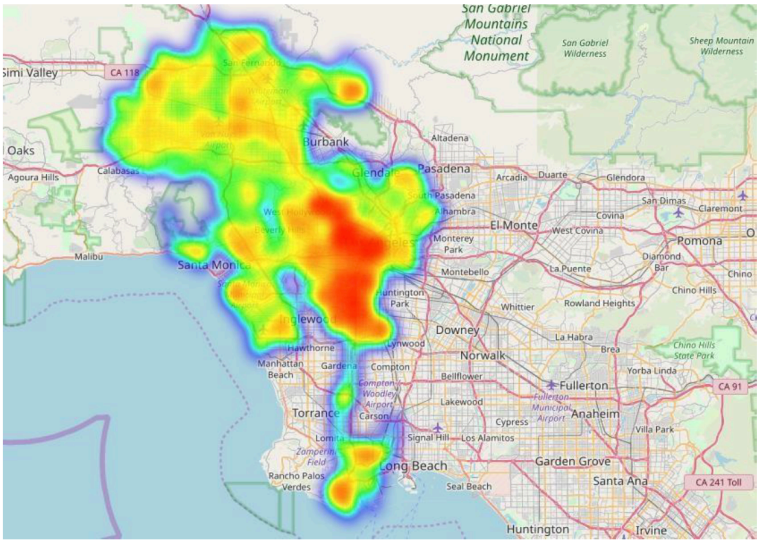


FIGURE 5, HEAT MAP OF ALL THE COLLISIONS FROM 2017 TO 2018



7, HEAT MAP OF CRASHES BEFORE SUNRISE AND AFTER SUNSET

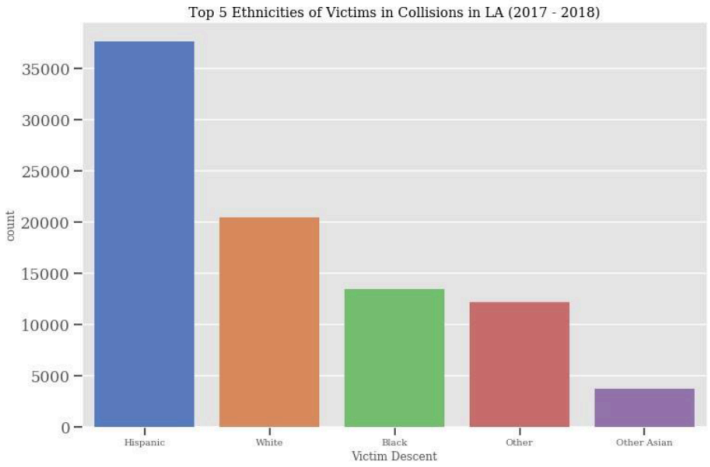
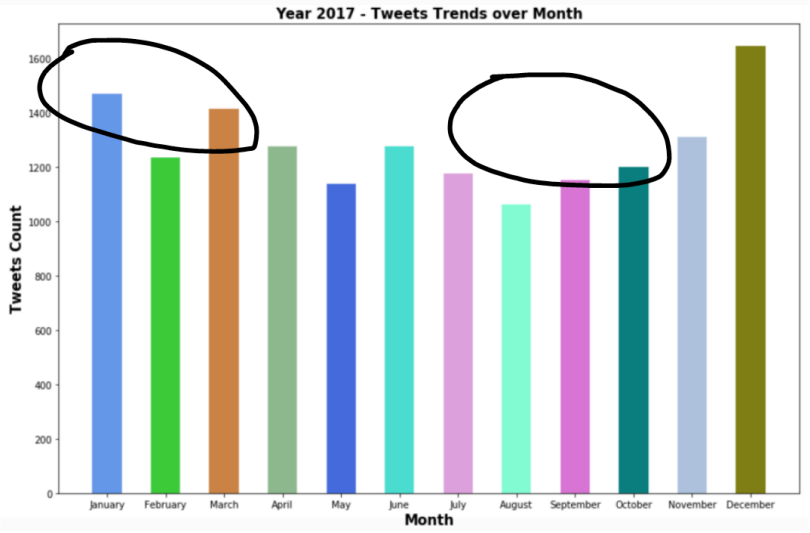
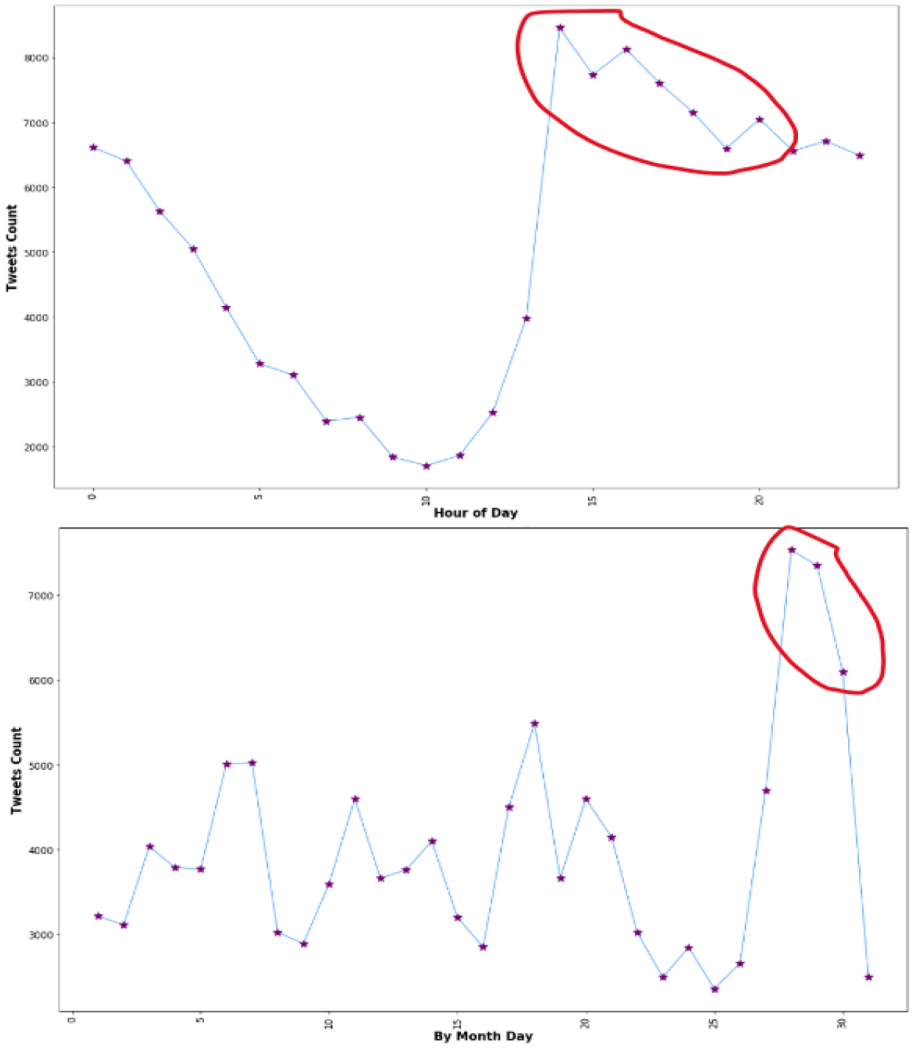
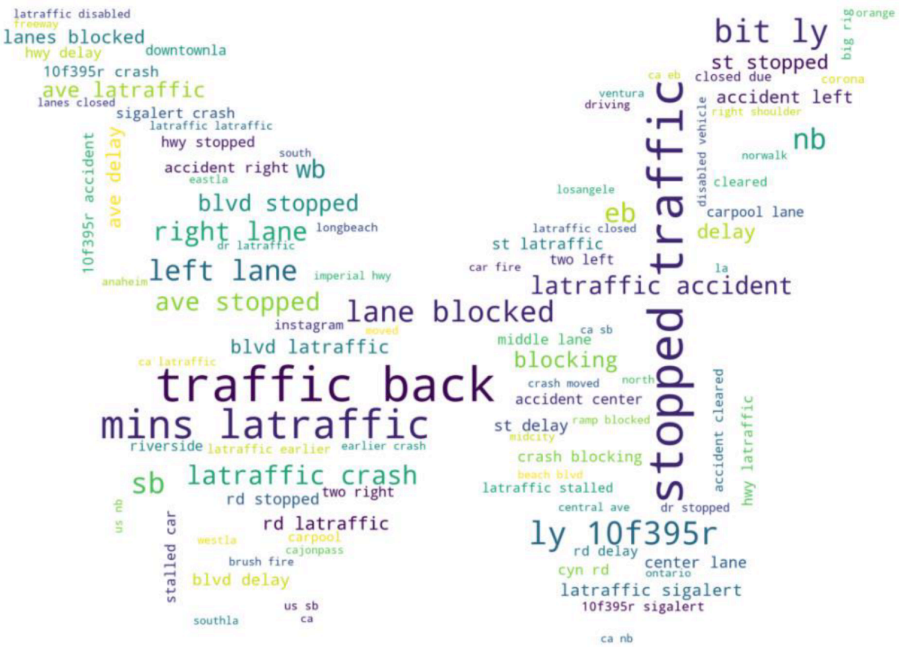


FIGURE 19, TOP 5 ETHNICITIES IN COLLISIONS




Tweets Count
Hour of Day
Tweets Count
By Month Day



CONCLUSIONS AND RECOMMENDATIONS

Los Angeles is a large and growing city which continues to attract more residents and vehicles. With the increase in the number of cars on the road comes additional traffic collisions and congestion. The analysis of the 2017-2018 data provides some interesting insights that can be used by a variety of people who are interested in LA traffic.

Below is a summary of our observations from this analysis:

1. What address/cross street combinations had the most collisions?
2. What are the most dangerous intersections?
    a. SEPULVEDA BL & SHERMAN WY
    b. NORDHOFF ST & TAMPA AV
    c. WHITSETT AV & SHERMAN WY
    d. RODEO RD & LA BREA AV
3. What are the most common collision areas in Los Angeles?
    a. 77th Street Area
    b. Council Districts 12/13/14
    c. Generally in the heart of LA
4. What are the best/worst times of the day for accidents? Best/worst month?
    a. Friday has the highest frequency of collisions.
    b. Sunday has the fewest amount of collisions.
    c. March and October have the highest number of collisions.
    d. The hours between 12PM to 5PM have the highest frequencies of collisions.
5. What patterns occur due to the amount of natural sunlight?
    a. There appears to be a significantly less concentration of accidents in Northern LA during the night time. This may be due to less traffic by the airports.
    b. The highest frequency of accidents is on Monday-Friday between 45pm.
6. What is the demographic makeup of victims in collisions?
    a. Men are more likely to be in an accident compared to women.
    b. Frequency of collisions is proportional to race/ethnicity.
    c. Age 30 has the highest number of collisions.
    d. 34.08% of accident victims have a median income between $40-49K.
7. Do certain temperatures or weather play a factor?
    a. When it rains during evening rush hour traffic, there is an increase in the number of collisions around 5pm.
    b. The area near the Hawthorne airport appears to have a higher proportion of weather-related accidents.

6. Market Basket Analysis

- Description

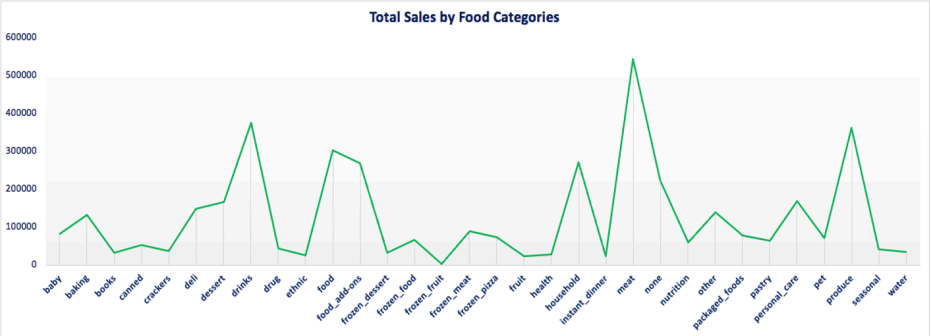
  Given our customer base of roughly 2,500 households with ~45,000 transactions over 2 years of shopping at a retailer, our goal is to segment and profile our customers based on factors such as income, shopping behavior, and demographics; then use this data to identify frequently paired items to develop a direct mail coupon campaign

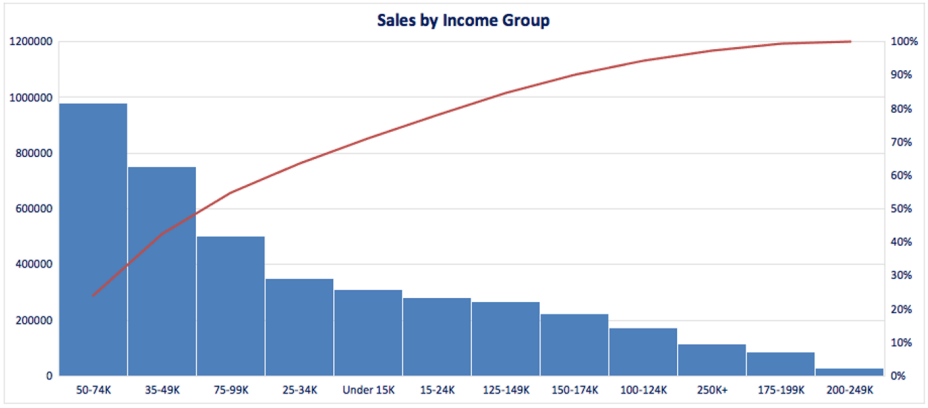
  1. Who are potential coupon users?
  2. What's the relationship between price, quantity, and coupon value?
  3. Which are the most frequent items sold?
  4. What's the optimal coupon value for selected most frequent items?

- Technology
  - Python
- Source Code
  - https://github.com/sathish-rajendiran/mar653/tree/main/Project/Final_Submission

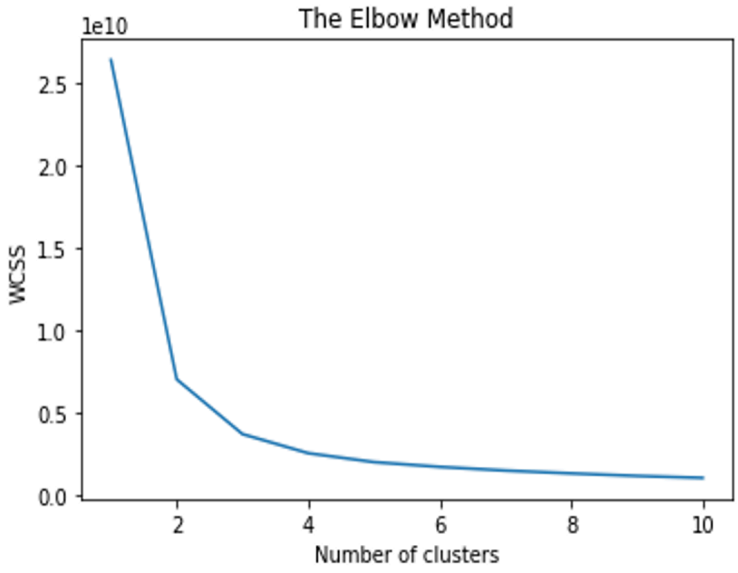
- Highlights

household_key	identitifies unique households
year	Year =1 or Year = 2
category	Product Category. Example "baking", "deli", "meat", etc
quantity	Number of products purchased during the trip
sales value	Amount of dollars retailer recieves from sale
price	Unit price
coupon_discount	Discount applied due to manufacturer coupon
dh_coupon_redeemed	Number of Store coupons redeemed
value of dh coupon	Value of store coupon
age_desc	Estimated age range (19-24, 25-34, 35-44, 45-54, 55-64, 65+)
marital_status_code	Marital status (A- Married, B - Single, U- Unknown)
income_desc	Household income (100-124k, 125-149k, 150-174k,..)
homeowner_desc	Homeowner, renter, probable owner, probable renter, unknown
hh_comp_desc	household composition (1 Adult 2 kids, 2 adults kids, single family, single male, unknown)
household_size_desc	size of household upto 5+
kid_category_desc	Number of children present (upto 3)





The Elbow Method for K means clustering was used to determine an optimal number of clusters of 3



CUSTOMER SEGMENTS AND PROFILE

Buys everything	Buys products for baby and children, and fresh food	Buys fresh food and household items mainly
Least likely to use coupons	Most likely to use coupons	Marginally more likely to use coupons
Age distribution general	Age Group 19-24 , 25-34 and few of 35-54	Age Group 35-54
Marital Status general	Mostly Married with large families	Mostly Married

Performed Apriori Rule Association

- Confidence (conf): how often the rule has been found to be true
- Support (supp): how frequently the itemset appears in the dataset
- Lift (lift): ratio of observed support to expected if x & y were independent
- Conviction (conv): ratio of expected frequency that x occurs without y

Segment (Cluster) 2:

- drinks, frozen_pizza -> meat
- baking, food -> food_add-ons
- dessert, packaged_foods -> meat

Cluster	Cart1 ->	Cart2	conf	supp	lift	conv
2	drinks, frozen_pizza	meat	0.943	0.085	1.265	4.454
2	baking, food	food_add-ons	0.928	0.217	1.417	4.803
2	dessert, packaged_foods	meat	0.928	0.083	1.244	3.512

- Average yearly sales for drinks, frozen pizza, baking, food, dessert, packaged foods, meat, and food add-ons yielded $1,062.34 in gross revenue.
- In an ideal scenario with 100% participation, sales for this item averaged across the year with the modified discounts yields an estimated gross revenue of $6,803.23
- This represents an ideal lift of $5,740.90, a **137%** increase in sales.

- Customers identified in Cluster 2 are the target group for this promotion.
  - These customers tend to have smaller families, are married between 35 and 54, and use coupons slightly more often than the average customer.
- If customers buy:
  - a drink and frozen pizza, meat is 14% off.
  - a baking item and any food item, add-ons are 7% off.
  - a dessert and a packaged food, meat is 14% off.
- Promotions can be combined for redemption against one item for each coupon.
- Run campaign for one month to determine if expected liability is met, and targeted customer group is responding to the promotion.

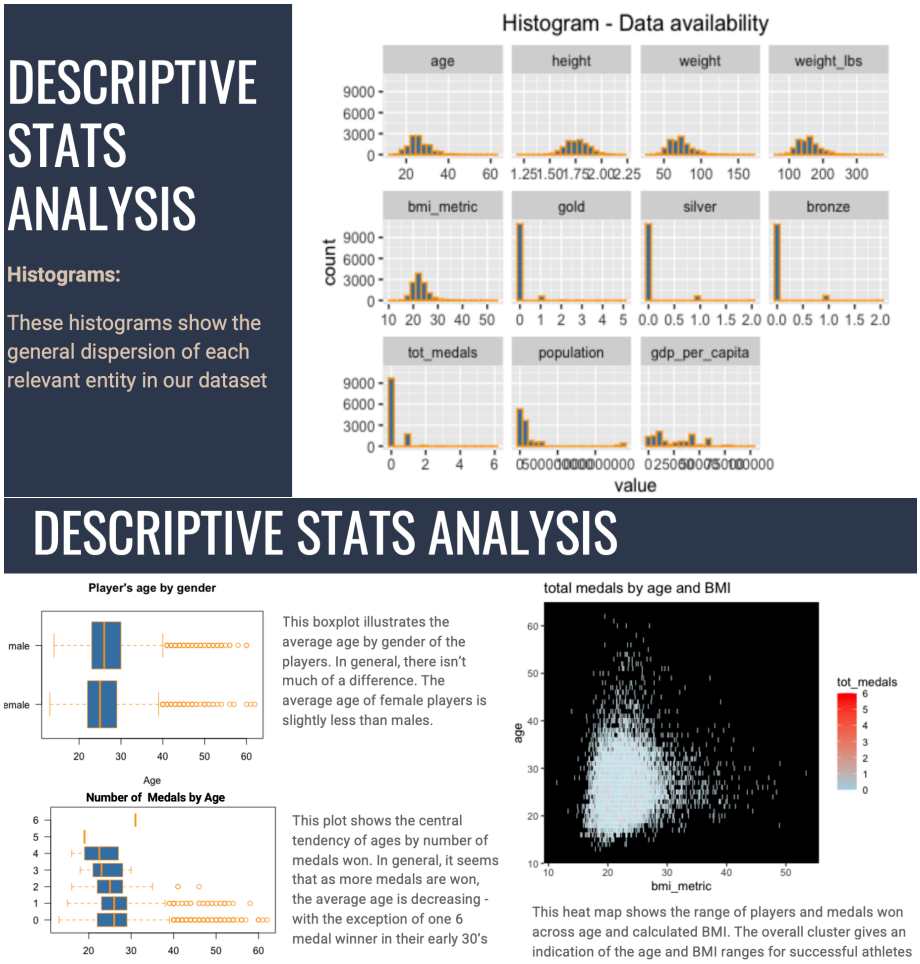
7. Predicting Athletes Sponsorship

- Description

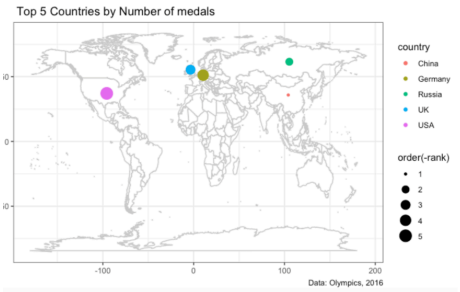
  Predicting Athlete sponsorship with 2016 Rio Olympic games data

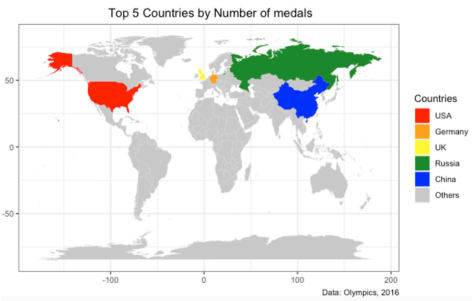
  1. Does an athlete's estimated BMI affect their chances of winning a medal?
  2. Is there a relationship between country and number of medals won?
  3. Which athletes are most likely to be sponsored?

- Technology
  - R
- Source Code
  - https://github.com/sathish-rajendiran/ist687/tree/main/Final_Project
- Highlights



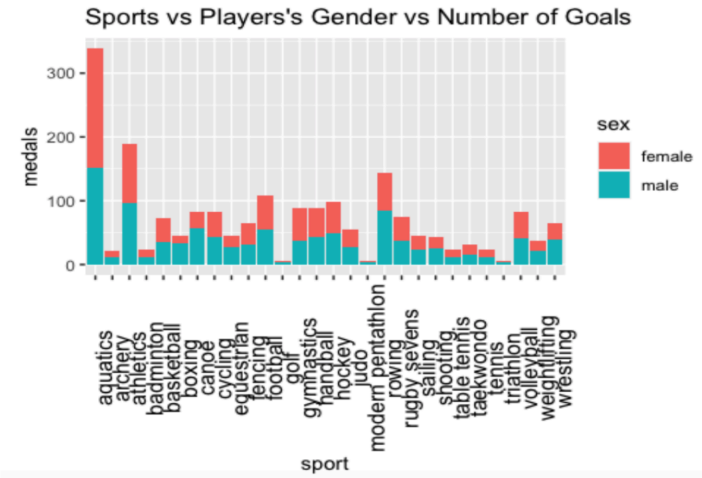
WORLD MAP: MEDALS BY TOP COUNTRIES





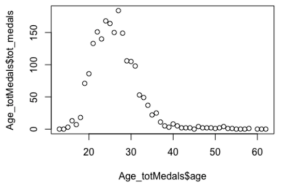
These maps shows the top countries by number of medals won with the USA having the most

BAR CHART: MEDALS BY SPORT



The bar chart to the right shows the breakdown of medals by sport, and is further broken down into gender. Aquatics won the most medals, and more females won medals than males.

LINEAR REGRESSION*

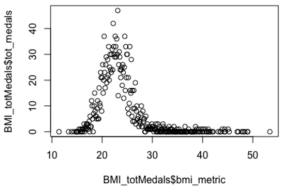


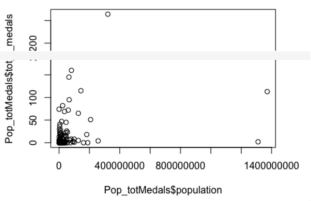
R-squared: 0.2647
p-value: 0.0000926

Assuming an alpha of 0.05, the p-value of 0.0371 indicates that age does have a correlation to total medals won

R-squared: 0.1904 p-value: 0.00000000000004687

Assuming an alpha of 0.05, the p-value of 0^-14 indicates that BMI does have a correlation to total medals won





R-squared: 0.09269
p-value: 0.00002755

Assuming an alpha of 0.05, the p-value of 0^-6 indicates that population does have a correlation to total medals won. Of note is that the p-value for BMI & population is smaller than age, so we will run the regression for age, BMI & population as out independent variables and analyze the output.

* Likely better fit with some other regression model, however due to the scope of the project we looked instead at where we could see significant correlation to total medals

PREDICTIVE ANALYSIS

Medal prediction equation using multivariable regression model coefficients:

Test 1

Predict total medals for a Male with BMI = 22, is 20 years old.

Y = 0.0722 + 0.004246x1 - 0.0005232x2 + 0.00000x3 + 0.024199x4

Y = 0.829 so we can assume that this person will win one medals

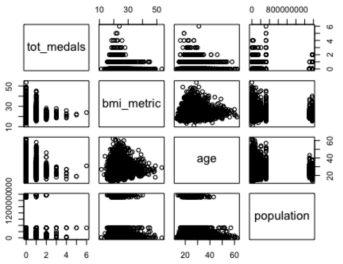
Test 2

Predict total medals for a Male with BMI = 30, is 45 years old.

Y = 0.200 medals so we can predict that this person is unlikely to win any medals

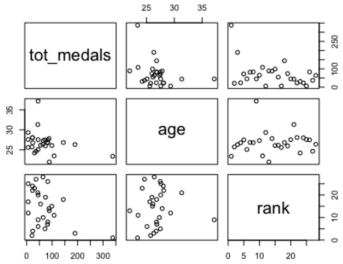
```
Call:
lm(formula = tot_medals ~ bmi_metric + age + population.x + MF,
    data = athletes)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3579 -0.1802 -0.1672 -0.1529  5.8129

Coefficients:
                 Estimate        Std. Error t value        Pr(>|t|)
(Intercept)   0.07224584434923  0.03297260665880   2.191        0.028465 *
bmi_metric    0.00424604702456  0.00124757756500   3.403        0.000668 ***
age          -0.00052328675194  0.00052918302954  -0.989        0.322753
population.x  0.00000000009197  0.00000000001420   6.477 0.0000000000975 ***
MF1           0.02419932284825  0.00844913847189   2.864        0.004189 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4287 on 11757 degrees of freedom
  (83 observations deleted due to missingness)
Multiple R-squared:  0.005002,  Adjusted R-squared:  0.004664
F-statistic: 14.78 on 4 and 11757 DF,  p-value: 0.000000000004788
```

MULTILINEAR REGRESSION



From the regression output, population and BMI have a p-value below our alpha of 0.05 indicating they have a stronger correlation to total medals won than age

Predicting total medals with age and ranked sports



TOP ATHLETES BY BREAKDOWN OF MEDAL TYPE

The athletes above were awarded more than 2 medals in the 2016 games - a majority of these athletes are swimmers and many compete in track & field events or gymnastics.

8. Zillow Pricing Analysis

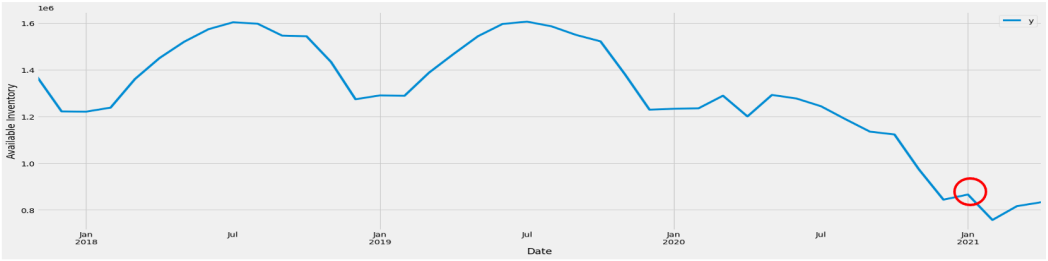
- The Problem
  - Due to COVID-19 changing social and professional constructs, people have been in a frenzy to buy homes (having saved money from inability to go outside or participate in events). From the news, online publications, and personal experiences, we can see that this has had a significant impact on the overall housing market, from both the buyer's perspective and the seller's perspective.
- Hypotheses
  - Buyers who are financially capable of purchasing a home, will be more likely to get the lowest interest rates for next few years than those who do not purchase now. However, buyers will experience highest home prices and severe supply & demand shock due to COVID-19.
  - Sellers who can sell their home between March 2020 – EOY 2021 will be more likely to retrieve the highest home sale price for the next few years than those who do not sell now.
- Data
  - Zillow Data (Home Inventory, Median List Price, Home Value Index)
  - FRED Data (Emp. to Pop. Ratio, Unemp. Rate, 15YR/30YR Mortgage Rate Average – all 2001 – 2021)
  - Opendatasoft Data (Latitude and Longitude by US Zip Code)
- Observation

  Based on data used, we can conclude that COVID-19 has impacted the housing market by
  - Home Inventories have reached their **lowest** point in the last 3 years
  - Home Median List Prices have reached their **highest** point in the last 3 years
  - Home Values have reached their **highest** point in the last 24 years
  - Employment to Population Ratio reached **lowest** point in the last 20 years
  - Unemployment Rate reached **highest** point in the last 20 years
  - 15YR and 30YR Fixed Mortgage Rates reached **lowest** point in 20 years
- Modeling techniques used:
  - Autocorrelation, Partial Autocorrelation
  - SARIMAX models focusing on CA, NC, and VA
    - Auto ARIMA & Grid Search Framework
  - Prophet (Entire U.S.)
- Based on our observation and analysis, we would recommend the following:
  - Should buyers purchase now?
    - If you can afford home at current MLP/Home Value, buy soon while Interest Rates are at 20-year low for 15YR and 30YR Fixed Mortgages
  - Should sellers list now?
    - If you can sell your home, consider selling in this market while Median List Price and Home Values are at an all-time high, and Home Inventory is at lowest point in last 3 years and expected to continue to drop.
- If we could include additional data or conduct additional analysis
  - New build/construction home data - impact on home inventory / supply & demand
  - Stock prices - 401k / liquidation for down payment
  - Leverage different modeling techniques

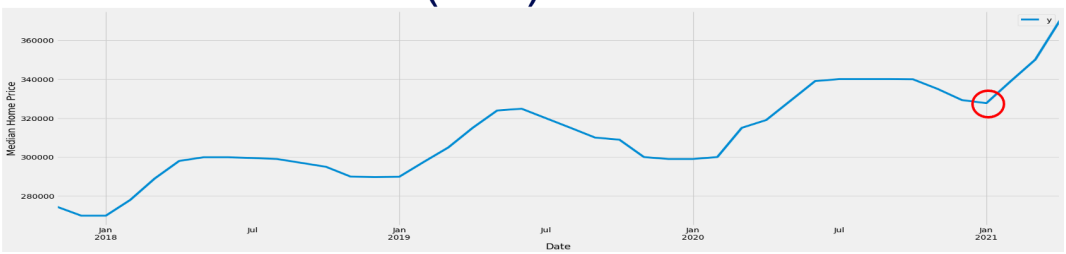
- Source Code
  - https://github.com/sathish-rajendiran/ist718/tree/main/Final_Project

Exploratory Analysis – Zillow Data

Available Inventory (U.S.) 2017 - 2021

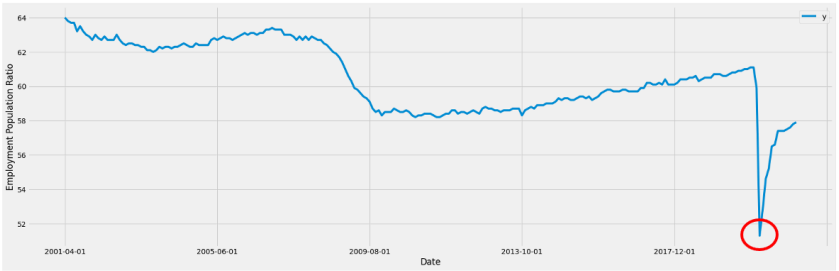


Median Home List Price (U.S.) 2017 - 2021

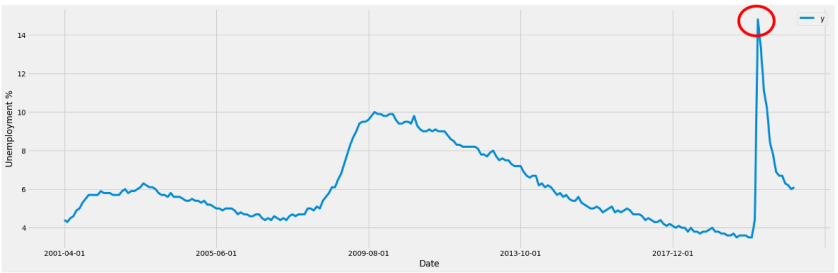


Exploratory Analysis – FRED Data

Employment to Population Ratio 2001 - 2021



Unemployment Rate 2001 - 2021

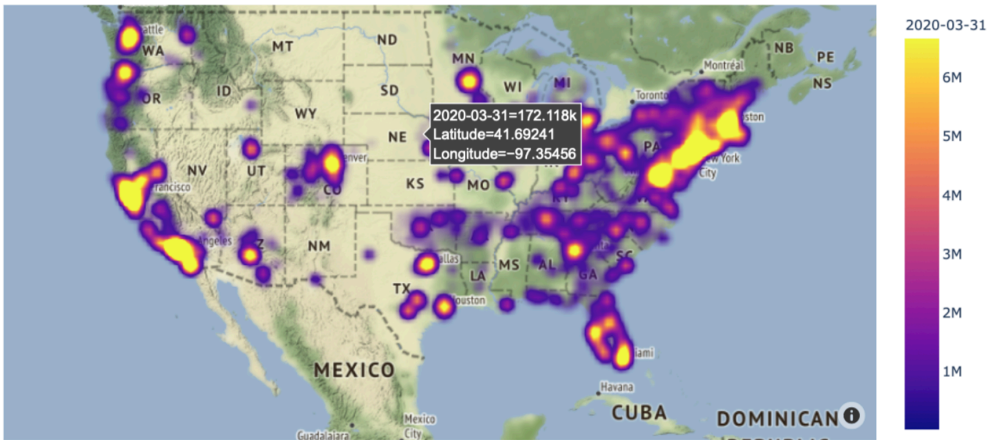


Exploratory Analysis – FRED Data

15YR Fixed Mortgage Rate Average 2001 - 2021

30YR Fixed Mortgage Rate Average 2001 - 2021

Interactive Map Visualization



Zillow Data Analysis Dashboard

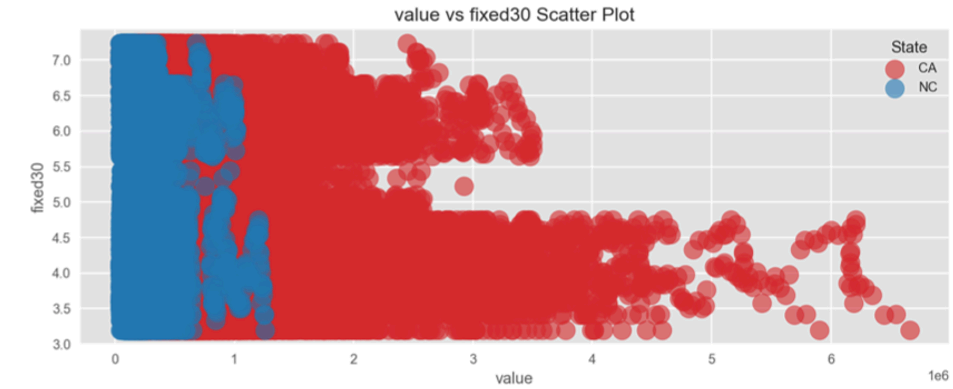
Please feel free to try various combinations of parameters to analyze results from different angles

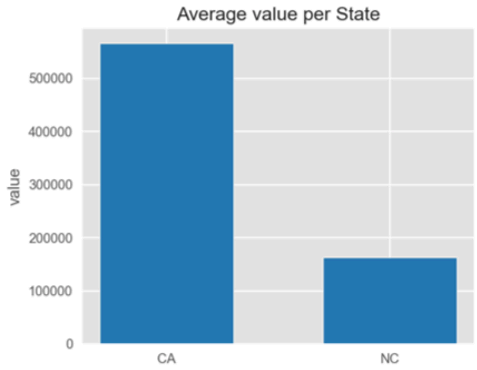
Scatter Plot Feature 1: value

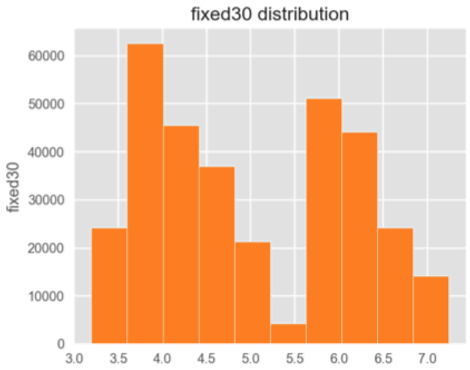
Scatter Plot Feature 2: fixed30

Bar Chart Feature: value

Histogram Feature: fixed30

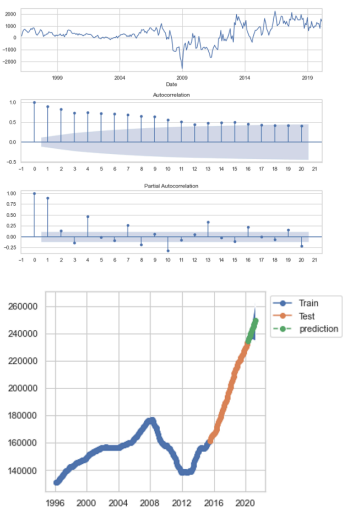






SARIMAX Results – Charlotte, North Carolina

Zipcode: 28269



SARIMAX Results

Dep. Variable:	28269	**No. Observations:**	291
Model:	SARIMAX(1, 2, 1)	**Log Likelihood**	-2092.872
Date:	Thu, 17 Jun 2021	**AIC**	4191.744
Time:	00:45:33	**BIC**	4202.743
Sample:	01-31-1996	**HQIC**	4196.151
	- 03-31-2020		
Covariance Type:	opg		

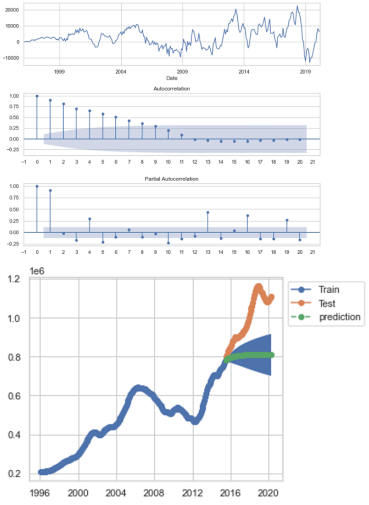
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6594	0.259	2.550	0.011	0.153	1.166
ma.L1	-0.7153	0.256	-2.792	0.005	-1.217	-0.213
sigma2	1.118e+05	5969.120	18.723	0.000	1e+05	1.23e+05

Ljung-Box (L1) (Q):	6.49	**Jarque-Bera (JB):**	130.37
Prob(Q):	0.01	**Prob(JB):**	0.00
Heteroskedasticity (H):	11.55	**Skew:**	0.37
Prob(H) (two-sided):	0.00	**Kurtosis:**	6.21

Official Zip Code Name	Charlotte
Zip Code State	North Carolina
Zip Code Type	Non-Unique
Primary County:	Mecklenburg
Secondary County:	Cabarrus
Area Code	704 / 980
Current Population:	71048
Racial Majority:	White 37.78%
Public School Racial Majority:	Black 59.9%
Unemployment Rate:	5.9%
Median Household Income	$79094
Average Adjusted Gross Income	$56960
School Test Performance:	Average
Average Commute Time	24.9 Minutes
Time Zone:	Eastern Daylight Time
Elevation Range	692 - 804 ft.
Area	31 Sqm.
Coordinates(Y,X)	35.33744700, -80.80241800

SARIMAX Results – Fremont, California

Zipcode: 94536



SARIMAX Results

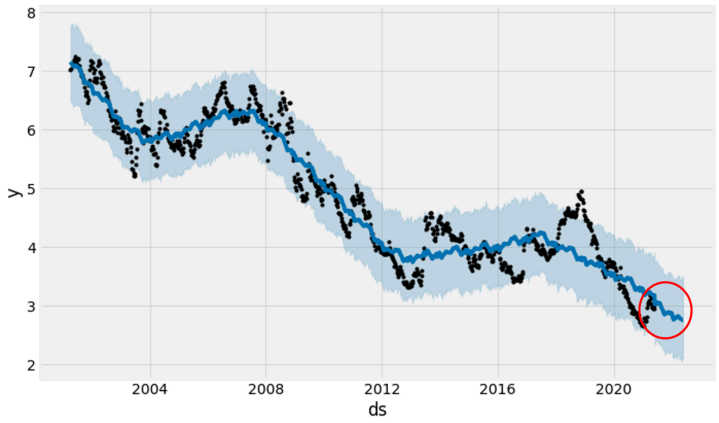
Dep. Variable:	94536	**No. Observations:**	291
Model:	SARIMAX(1, 1, 1)	**Log Likelihood**	-2907.002
Date:	Thu, 17 Jun 2021	**AIC**	5820.004
Time:	00:38:22	**BIC**	5831.014
Sample:	01-31-1996	**HQIC**	5824.415
	- 03-31-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8614	0.021	41.227	0.000	0.820	0.902
ma.L1	-0.7523	0.026	-29.455	0.000	-0.802	-0.702
sigma2	2.94e+07	8.29e-11	3.55e+17	0.000	2.94e+07	2.94e+07

Ljung-Box (L1) (Q):	189.38	**Jarque-Bera (JB):**	275.40
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	2.05	**Skew:**	-0.88
Prob(H) (two-sided):	0.00	**Kurtosis:**	7.44

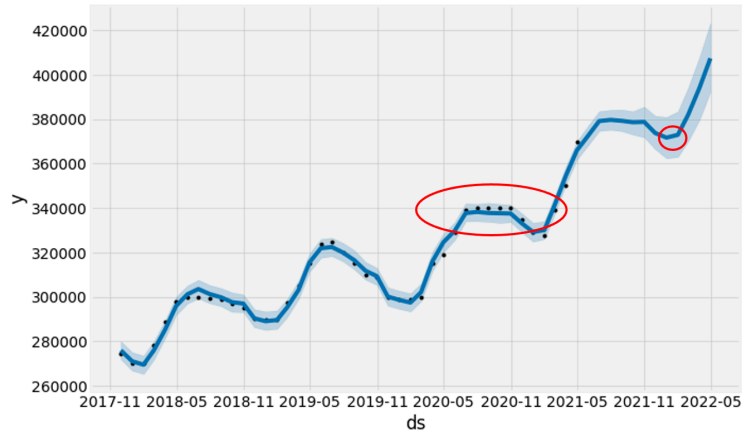
Official Zip Code Name	Fremont
Zip Code State	California
Zip Code Type	Non-Unique
Primary County:	Alameda
Area Code	341 / 510
Current Population:	68790
Racial Majority:	White 32.07%
Public School Racial Majority:	Asian 50.2%
Unemployment Rate:	5.4%
Median Household Income	$86651
Average Adjusted Gross Income	$123680
School Test Performance:	Above Average
Average Commute Time	29.3 Minutes
Time Zone:	Pacific Daylight Time
Elevation Range	121 - 1001 ft.
Area	14 Sqm.
Coordinates(Y,X)	37.57097500, -121.98795300

Prophet Model, 30-year Interest Rates



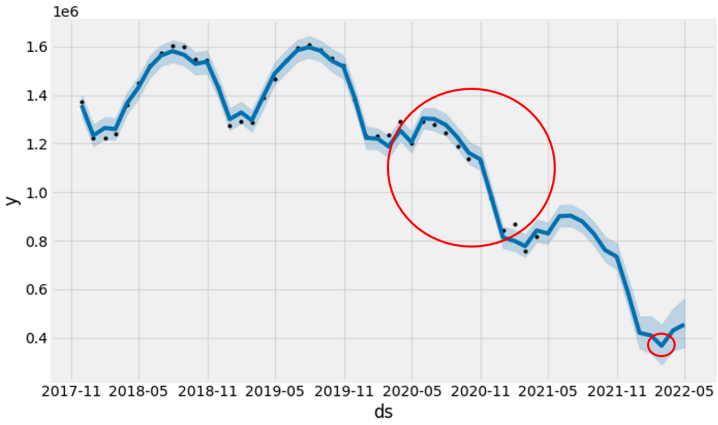
- Interest rates are projected to continue to decline for a 30-year fixed rate loan.

Prophet Model, U.S. Median Home List Price



- Based on the prediction, the rest of 2021 appears to have plateaued.
- A brief decline in the median home price is expected around February 2022.

Prophet Model, U.S. Inventory



- Inventory is expected to continue to decline until an expected to slightly rebound about February 2022.

Summary

Overall, this program has made me to look at data problems more efficiently. In addition,

- Effective storytelling
- Building data pipelines using R & Python languages.
- Processing tweets, clickstream and social network data using NoSQL technologies
- Process improvements using DMAIC principles (L & A)
- Data Analytics, Text Mining, Machine Learning, Statistic Modeling, Predictive, Prescriptive Modeling and Natural Language Processing (NLP)
- Hands-on experience on Python libraries like NumPy, Pandas, Matplotlib, Seaborn, NLTK, Sci-kit learning, SciPy, TensorFlow, Keras & PyTorch
- Experience in Text Analytics, developing different Statistical Machine Learning, Deep Learning, Data Mining solutions to various business problems and generating data visualizations using R, Python and creating dashboards using tools like Tableau & Qlikview
- Natural Language Understanding (Sentiment Analysis, Custom Analyzers, Entity Analysis, Word embedding)
- Natural Language Processing-NLP (LSA, LDA, TF-IDF, Markov Models, Tokenizers, Analyzers, POS tagging)
- Unsupervised techniques (PCA, K-Means and Hierarchical Clustering)
- Dimensionality reduction techniques (PCA)
- Non-Parametric Fast Learning ML Algorithms (Decision Trees, Random Forest, Gradient Boosting (Xgboost, Light Gradient Boosting), SVM)
- Deep Learning Models (Neural Network: CNN, RNN), Deep Learning Frameworks (Tensor Flow, Keras)

Next Steps

- Advanced Research/enrolling in PhD program
- Participating in Kaggle Competitions
- Student Ambassador - Helping prospects and students to excel in this program

References

- Textbook and Readings
  - **Discovering Statistics** by Daniel T. Larose - 3rd edition
  - **Understanding Variation - The Key to Managing Chaos**, 2nd edition By Donald J. Wheeler; SPC Press
  - Hoffer, J. A, Ramesh, V., & Topi, H. (2016). **Modern database management** (12th ed.). New York, NY: Pearson.
  - **Managerial Analytics: An Applied Guide to Principles, Methods, Tools, and Best Practices**, December 2013, 1st Edition, Watson, and Nelson
  - **A Practitioner's Guide to Business Analytics: Using Data Analysis Tools to Improve Your Organization's Decision Making and Strategy,** 2013, Bartlett
  - **Introduction to Data Science (2017**), by Jeffrey S. Saltz & Jeffrey M. Stanton.
  - Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) **Introduction to Data Mining.**
  - Tom Mitchell (1997) **Machine Learning**
  - Brett Lantz (2015) **Machine Learning with R (second edition).**
  - Stanton (2017), **Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R**
  - Bird, S., Klein, E., & Loper, E. **Natural language processing with Python**
  - Jurafsky, D., & Martin, J. H. **Speech, and language processing** (3rd ed. draft).
  - Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (2005) **Introduction to Data Mining**
  - Venkatesan, R., Farris, P. and Wilcox, R. (2019). **Cutting Edge Marketing Analytics: Real World Cases and Data Sets for Hands on Learning**
  - Yau, N. (2011). **Visualize this: The Flowing Data guide to design, visualization, and statistics**. Wiley Publishing.
  - Yau, N. (2013). **Data points: Visualization that means something**. Wiley Publishing.
  - **A Smarter Way to Learn Python** by Mark Myers (available on Amazon)
  - **Python for Everybody:** https://www.py4e.com/book (Python version)
  - Miller, Thomas W., **Modeling Techniques in Predictive Analytics with Python and R**, Pearson, 2015.
  - Goodfellow, Ian, Yoshua Bengio, and Aaron Courville, **Deep Learning (DL)**, MIT Press, 2016
  - James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, **An Introduction to Statistical Learning with Applications in R,** Springer, 2013

- Source Repository

  https://github.com/sathish-rajendiran?tab=repositories

- Resume

  https://www.linkedin.com/in/sathish-kumar-rajendiran-2963599/