CAPITAL ONE

DATA SCIENTIST CHALLENGE - Pranesh Narayanan

Mail: - pxn151730@utdallas.edu Mobile: - +1-469-888-9234

Objective – Predict the target variable (Numerical) using the 254 features from the 5000 records

PART 1:- Model Building on Synthetic Data

Implemented the model using R

Step 1:-

Loaded the provided codetest_train text file into R Studio as codetest_train(R dataset)

Initial Understanding:-

1. Training set consists of 4 categorical variables –
   - f_61(Levels: a,b,c,d,e)
   - f_121(Levels: A,B,C,D,E)
   - f_215(Levels: Red, Orange, Yellow, Blue)
   - f_237(Levels: Mexico, Canada, USA)
2. All the numerical data provided were in the standardized form with mean close to 0 and SD close to 1
3. The null values in every explanatory variables were between 100-200; hence cannot run regression without imputation
4. The given target variable has high variance and hence the accuracy will take a hit

Step 2:-

The imported dataset has been imputed using MICE package - Predictive Mean Matching method and m = 3. The best imputed model can be chosen after the initial regression.

*Note: - Target variable has been removed to eliminate the imputation bias

Step 3:-

Linear Regression was performed to understand the significance of input variables and to choose the best imputed dataset. The first dataset was chosen using Adjusted Rsq measure.

-Adjusted Rsq value (Since, too many variable are there, Adj Rsq is a better model fit value than Rsq)

Step 4:-

Cbind was used to merge the target variable to the input attributes for further analysis.

The Merged-Imputed Dataset – **Regdata**

***Note**:- PCA and stepwise(forward and backward) have not yielded any satisfactory results. Hence have not reduced or removed features

Step 5:-

Partitioned the Regdata dataset into training (4000 records - trainData) and Validation (1000 records – testData). Partitioning will help determine the best model (the least MSE value)

Step 6:-

Have implemented the model across 6 regression techniques

- Linear Regression
- Support Vector Machine Regression
- Neural Nets
- Random Forest Regression
- Gradient Boosted Regression

SVM Regression has consistently provided minimum MSE and hence will be used as the prediction model for the given dataset

Step 7:-

Have loaded the provided test dataset into R Studio - codetest_test and performed multiple imputation using MICE package. 3 imputed test datasets are obtained.

Step 8:-

SVM regression is used to model the 3 test datasets and predicted the target value are obtained.

The target values obtained have been averaged to achieve a value closer to the Expected Value

The prediction averaged dataset is exported to text – **predictonmain.txt**

Summary: Statistics-

Provided target [Mean – **1.143878**, Variance – **27.66651**]

Predicted target [Mean – **1.035195**, Variance – **12.39833**]

CAPITAL ONE

DATA SCIENTIST CHALLENGE - Pranesh Narayanan

Mail: - pxn151730@utdallas.edu Mobile: - +1-469-888-9234

Objective –Identify a new and refreshing insight from Baby Names dataset

PART 2:- Baby Names

Have utilized Tableau 9.2 to work on this question.

A. Descriptive Statistics

Merged the text files using – “copy /b*.txt newfile.txt” in command prompt and have loaded the merged file –newfile.txt into the tableau

1. The input is provided using tab delimited text file.

   The issue with the dataset is that, only the first names have been provided in the input file (2-15 characters).
   Although, it captures the common naming features across states and years, the small deviation in spelling is treated as a different record entirely. (For eg – Sophia - Sofia and Alexander – Alexandr are treated as different names although they have the same pronunciation and might have the same origin)

   This creates a bias in the provided dataset and may not be very accurate.

2. Most popular name of all time – **James** [4,938,965(Total Number of Occurrences)]
3. Most gender ambiguous name:-
   Criteria Selected – Difference between the Male & Female occurrences is zero (Ratio = 1) and total occurrences across each Sex is maximum from the selected group

   2013:- **Nikita** (97 Total Occurrences [47 Male & 47 Female]

   1945:- **Maxie** (38 Total Occurrences [19 Male & 19 Female]

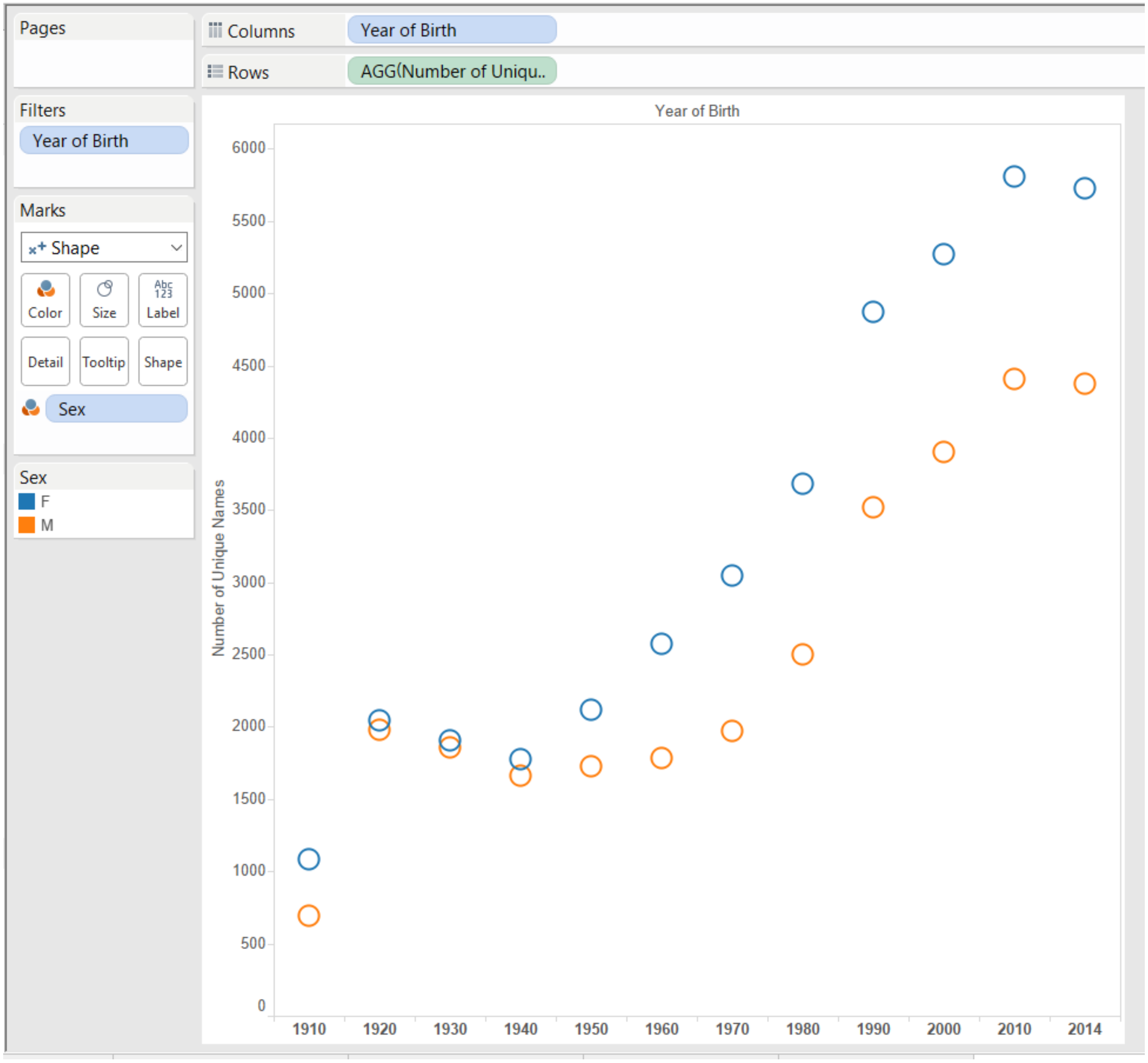
4. Comparing percentage Increase across each year from 1980 – 2014(Since was used in the question)

- **Aaliyah** has the highest percentage increase of **28,140% at 1994**
- **Alexandr** has the highest percentage decrease of -**98.37% at 2005**

5. Across the entire category, **Deneen** recorded the highest percentage increase of **31340%** at the year **1964**

B. Onward to Insight

1. Average Number of Unique Names Vs Year of Birth (Across Sex)



Pages

Columns: Year of Birth

Rows: AGG(Number of Uniqu..

Filters: Year of Birth

Marks: Shape

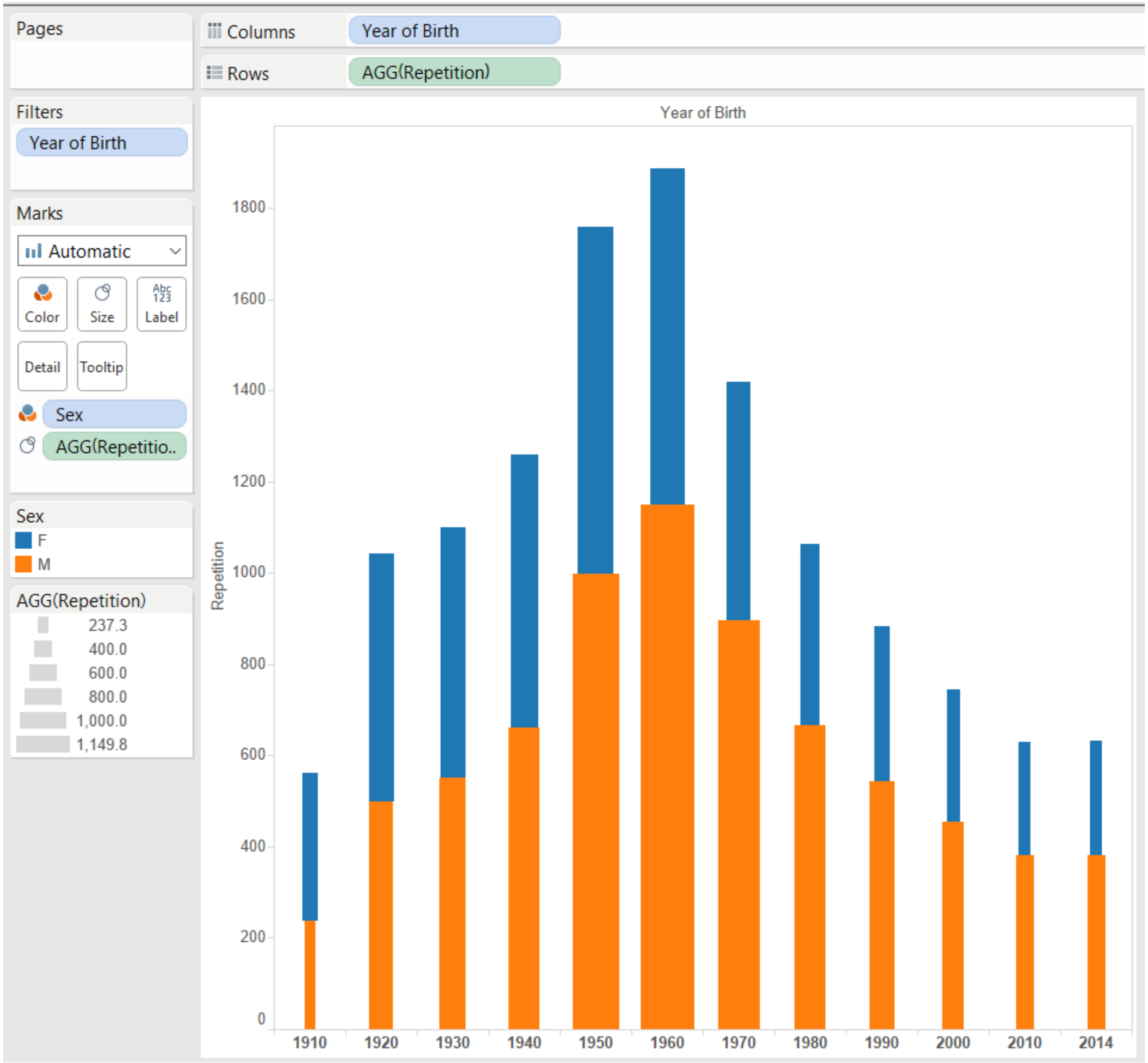
Sex

Insight:-

With the progress of time (Years), people have become more creative in naming and have come up with a lot of unique names. This can be attributed to the advent of Internet and cables. With so many programs and characters in action across multiple domains – Entertainment, Sports, Politics to name a few, and improvement of people s ability to stay updated on all the trends, astronomical number of choices for baby names have been made available

Also, it can be seen that, girls have a lot more unique names than guys. In general people tend to be more creative while naming girl babies than boys across the provided states

2. Understanding Repetition of Baby Names across Years



Columns: Year of Birth

Rows: AGG(Repetition)

Filters: Year of Birth

Marks: Automatic

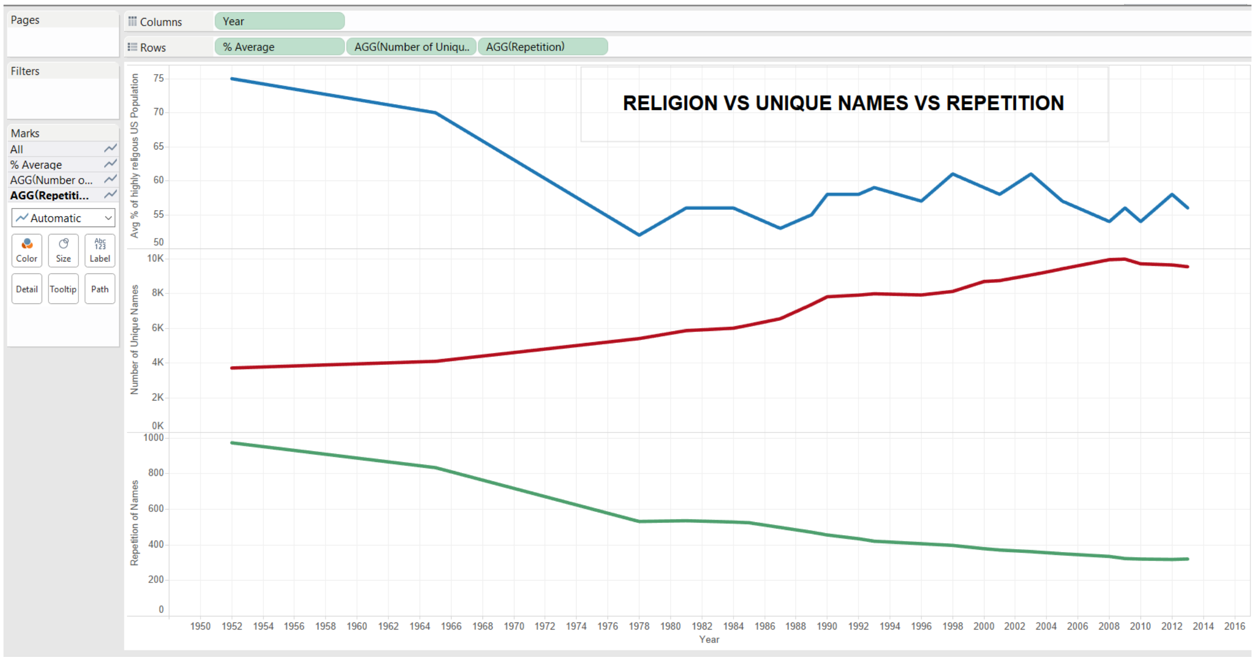
Sex

AGG(Repetitio..

Insight:-

In 1950s & 1960s, repetition of baby names were predominant. Although, this might imply that it may be related to the lack of communication and access to information on latest trends (since newspapers do not provide us with the choice of content) during those decades, another factor comes into play - **Religion**

Correlation between Religion, Repetition & Number of Unique names



Columns: Year

Rows: % Average | AGG(Number of Uniqu.. | AGG(Repetition)

Marks: All | % Average | AGG(Number o... | AGG(Repetiti...

Automatic

<u>Insight:-</u>

When we drill down on the probable names during the 1950s and 1960s , we find that the names such as Michael, Robert, James, Mary have the majority share and the repetition for these names are too high. These names are very common and can be attributed to the faith in Christianity religion. So, during these decades, names were often derived from Christian scriptures. Religion was the one of main modes of communication during this time.

Hence, when the data of % of US population who feel religion is very important in their life is plotted against time (Years), we can see a perfect correlation –

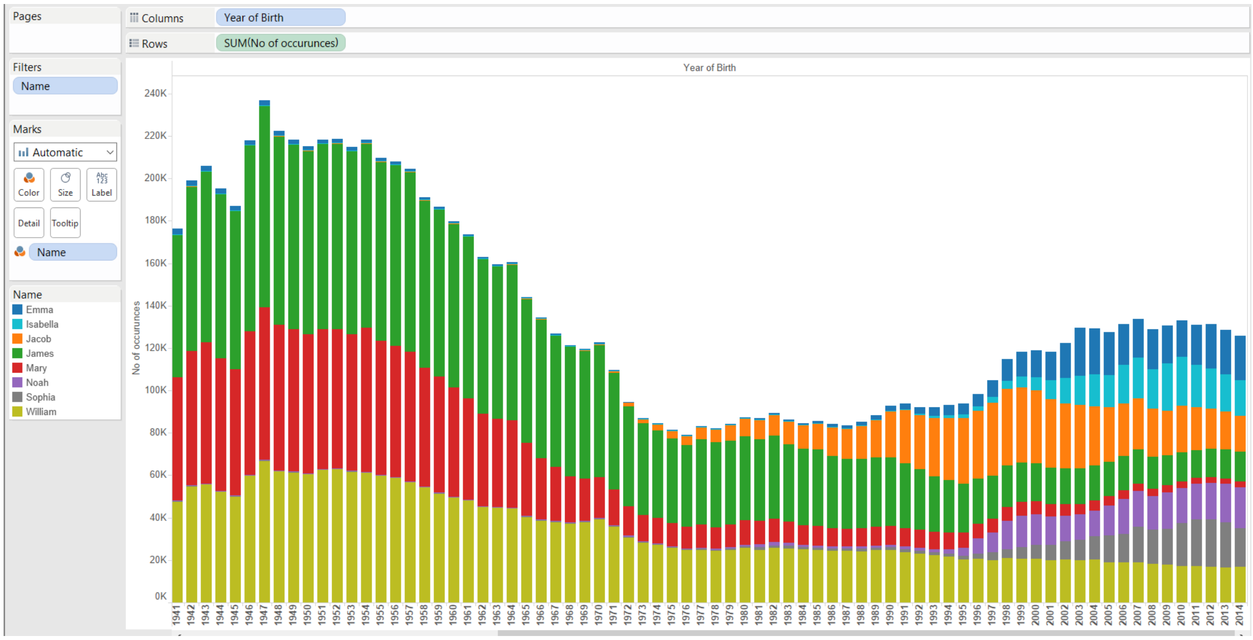
- As the average percentage of US population who feel religion is very important reduce, the number of unique names start increasing correspondingly (Negative Correlation)

- As the average percentage of US population who feel religion is very important reduce, the repetition of names start reducing correspondingly (Positive Correlation)

  This proves that religion has played a vital part in the naming criterion between 1940 - 1960, predominantly Christian (according to the dataset).

  Religion Data Obtained from – **http://www.gallup.com/poll/166613/four-report-attending-church-last-week.aspx**

3. Visualizing the distribution of Mary, James, William and some latest names


Columns: Year of Birth

Rows: SUM(No of occurunces)

Filters: Name

Marks: Automatic

Color: Name

The graph is color coded and the coding schema is provided in the chart.

It can see seen that, James, Mary and William have dominated in terms of number of occurrences throughout our Analysis time period (1910-2014). Although, the occurrence of Mary and William are slowly dampening over time, occurrence of James still seems to be significant.