BAN 5753

Team: Group #: 8 Mounica Mandapati Ashlesha Sampathi Nitesh Maruthi Venkat Ram Reddy

10/6/2019

SAS Symposium Initial Report



**SAS Symposium – Initial Report**

1. **Dataset: KickStarter**

Link: <https://data.world/rdowns26/kickstarter-campaigns>

This dataset has about 20 thousand observations with US and other countries (US observations being highest). We did some descriptive statistics on this data and the data looks good.

**Business statement:** Explain and predict the factors causing the success of a program. We plan to combine text analytics with other variables and analyze the failure and success of a program.

Understand the marketplace of Kickstarter including timing of campaigns posted, types of projects, location of campaigns, description of campaigns and more



1. **Dataset: YRBSS(Youth Risk Behavior Surveillance Survey)**

Link to the dataset: <https://www.cdc.gov/healthyyouth/data/yrbs/data.htm>

This data set is based on three sets of survey, school district and national. We wanted analyze lowest level data and started with district survey data but since fewer number of districts participated in this survey, we picked next higher-level dataset state-level survey.

YRBSS State level survey data set contains data from 1991 to 2017(2019 dataset will released at the end of year 2019).

Analyzing YRBSS state level data, it was observed that this dataset contains a million records with on average about 100 thousand records per survey year. Survey year are alternate years starting from 1991 and with time more and more states was included in the survey.

There are about 80 questions in the survey on safety(seat belt to drink and drive), nicotine consumption(all formats), drugs consumptions(all formats), Teen sex and safety, vegetable/fruits and physical exercise.

Initial analysis of the data between 2017 and 2015 was not a clear trend on one way or another but a decade data could give better picture by state based on different polices introduced by state.

Since trend analysis was not encouraging and based on previous experience we were told that passive sample data is not necessary reliable to give any good story.



So, we are analyzing 2017-year dataset only for these business problems:

1. What are the behavioral patterns in people before committing suicide?
2. Analysis of good habits (like drinking milk, eating fruits) Vs. bad habits (like drug consumption, suicidal behavior)
3. Co-occurrence of different health risk behaviors
4. **DataSet: Climate change – Forecast air quality in the US**

Link: <https://openaq.org>; <https://docs.openaq.org/>

<https://www.climate.gov> (temperature data)

Climate change is upon us and has been a burning issue from a long time. There is renewed interest in climate change recently due to changes in US policy and Greta Thunberg – the young activist who is making the world aware about the impact of pollution.

This dataset contains details of daily pollutant levels from all over the world – the pollutants recorded are PM2.5, PM10, ozone (O3), sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and black carbon (BC). The data is recorded daily at several locations all over the world in key cities.

**Business Statement:**

We can perform a time-series forecast on this data in order to forecast the deterioration of air quality in the US. Since data is available from all over the world, we are planning to limit our data to a particular location – Washington DC: the location of the conference. We are hoping to see a decline in the quality of air, and since the analysis will be done on the location where it will be presented, the audience will be able to relate to it better. This is an example of using data science methods to explain socio-political problems.

Since the data is recorded daily, we can download other datasets that can be merged such as daily temperature data (in link). We are presently searching for other datasets that can be combined with the air quality data. Likely target variables are temperature/pollutant levels. The dataset size would depend on the timeframe that is chosen for analysis (5/10/15 years).

The openAQ website provides direct download as well as an API to download the dataset. The cap is set to 2000 requests in a 5 minute window, but we did not see a cap on the response size per API call. The direct download is also available, but is slightly restrictive in terms of options.

Below is an example forecast that was done using this dataset on the city of New Delhi, India.

http://www.urbanemissions.info/delhi-air-quality-forecasts/daqi-hourly-timeseries/

1. **DataSet: Gender pay gap**

Link: https://data.gov.uk/dataset/54219db1-dd98-49d9-a383-a5978bb0aeb9/gender-pay-gap

This is the biggest dataset we could found on gender pay gap based on the size of the file available online. This file contains about 10,000 observation reported by various organization with mean data of salaries and bonus. We could make a descriptive study on the dataset but other have already working on it mostly using python. We are still not sure how to come up with a given predictive problem on this dataset.

1. **Dataset: DAWN, MHSS, NSDUH, MHSS and TEDS**

Link: <https://datafiles.samhsa.gov/info/browse-series-nid3453>

These datasets have drug use, mental health and admission in ER due to drug over dose in hospitals by state.

**Problem Statement:** Change in Substance abuse cases with opening up Marijuana.

We wanted to study change in substance abuse with change in marijuana laws in the state but due to lack of appropriate variable in the data and discussing with subject matter experts(ER doctors). We were not confident of a good story out of the dataset.

**Drug Abuse Warning Network –** SAMHSA is re-establishing the Drug Abuse Warning Network (DAWN), a nationwide public health surveillance system that will improve emergency department (ED) monitoring of substance use crises, including those related to opioids. Hospital participation will continue to be voluntary. Data abstraction will begin in mid-2019 with a group of 25 hospitals, and will grow to 50 hospitals in DAWN’s second year.

Comment: Problem with these dataset is that the data is available only until 2011 which is too old for the study.

# National Survey of Substance Abuse Treatment Services (N-SSATS)

The National Survey of Substance Abuse Treatment Services (N-SSATS) is an annual census designed to collect information from all facilities within the 50 States, the District of Columbia, and the U.S. territories, both public and private, that provide substance abuse treatment. N-SSATS provides the mechanism for quantifying the dynamic character and composition of the United States substance abuse treatment delivery system.

Comment: This data set can be used to study the impact of legalizing marijuana has on substance abuse. This data set has state information in their data, so we can use this data set to do state wise study based on when the state legalized marijuana and study the impact of substance abuse treatment

# National Survey on Drug Use and Health (NSDUH)

# The National Survey on Drug Use and Health (NSDUH) series, formerly titled National Household Survey on Drug Abuse, is a major source of statistical information on the use of illicit drugs, alcohol, and tobacco and on mental health issues among members of the U.S. civilian, non-institutional population aged 12 or older.

# Comment: The problem with this dataset is that a state variable is not included in the public-release datasets, so state based study cannot be done.

# Treatment Episode Data Set: Admissions (TEDS-A)

# The Treatment Episode Data Set -- Admissions (TEDS-A) is a national data system of annual admissions to substance abuse treatment facilities. State laws require substance abuse treatment programs to report their publically-funded admissions to the state. Some states collect only their publically-funded admissions while other states are able to collect their privately-funded admissions from facilities that receive public funding. States then report these data from their state administrative systems to SAMHSA. The resulting data system is referred to as TEDS-A. Thus, TEDS does not include all admissions to substance abuse treatment. It does, however, include that portion that would constitute the public burden for substance abuse treatment.

# Comment: The first three data sets are based on surveys; however, this dataset is based on actual admission. This might be a good dataset, but again not all states do a complete disclosure, so this might create a problem. State variable is available in coded format in the data set.

1. **Dataset: OKCupid dating app**

Link: <https://github.com/rudeboybert/JSE_OkCupid/blob/master/profiles.csv.zip>  
https://openpsych.net/files/papers/Kirkegaard\_2016g.pdf

This dataset is scraped from the dating app OKCupid from all over the world. It contains user profile data. The original dataset contained about 68,371 observations and 2,620 variables and was last updated in late 2015. There were privacy and T&C issues with this dataset, although the same was used in some papers for analysis. The linked dataset is a subset of the original, and contains 31 columns and nearly 28,000 observations. The observations include physical characteristics, lifestyle preference, education and text fields for a short introduction.

The original authors have made the scraping code public, but it needs to be modified due to changes in the website since the code was written.

We were looking into this dataset to design our own recommendation engine leveraging analytics, or come up with a custom model to get best 5 matches for a user.

Upon analysis, we decided to not use this data due to a lack of additional columns that can help us verify this information that can help us verify the results.

1. **Dataset: Marijuana consumption**

Link:<https://www.cdc.gov/brfss/questionnaires/pdf-ques/2018_BRFSS_English_Questionnaire.pdf>

This dataset is from Centers for Disease Control and Prevention (CDC) website. This survey is state-based Behavioral Risk Factor Surveillance System (BRFSS) a cross-sectional telephone survey that state health departments conduct monthly. BRFSS is used to collect prevalence data among adult U.S. residents regarding their risk behaviors and preventive health practices that can affect their health status.

This has information about the demographics, various diseases, drug consumption etc. summing to around 80 variables. We primarily concentrated on marijuana consumption state wise. We planned to understand the before and after effects of marijuana legalization in states. On preliminary analysis we found that it is not possible to do state wise analysis since not all variables are available for comparison.

Now we changed the business statement to observe the marijuana consumption at individual level.

**Business statement:** Explain and predict the causes leading to marijuana consumption.

On building naïve models, we came up with variables like age, tobacco consumption, smoking as the important variables explaining the marijuana consumption.

To make it more interesting we are planning to look into the trends in the way marijuana is consumed. Like, Are people eating more marijuana now compared to smoking it? And the factors leading to the change in the way marijuana is consumed.

1. **Dataset: UFO sightings**

Link: <https://data.world/timothyrenner/ufo-sightings>

We looked into this topic initially due to the hype around area-51 and UFO’s that has been renewed off-late by the raid area-51 movement. The dataset contains fields such as city, state, sighting time, UFO shape, report text, lat-long etc. Exploring the possibility of performing a time series analysis to look at trend and seasonality in depth – some basic EDA available online.

We decided to drop this topic as the dataset is not very detailed, would have been better if there were more variables as it is hard to build a detailed story around the results we may find beyond seasonality & cyclicity.

1. **Dataset: Cryptocurrency analysis**

Link: <https://www.quandl.com/data/BITFINEX/ETHUSD-ETH-USD-Exchange-Rate>

Etherium is a popular cryptocurrency, that is well known for the advanced blockchain network that it is built on. With the surge in interest and prices of bitcoin, etherium became the second most popular cryptocurrency. We looked at this data to see the possibility of predicting the next “big” event – sharp rise and fall, using the daily trading data. Quandl provides reliable daily trading data for a host of currencies such as BTC, ETH, USD and hundreds of others. Fields available in the dataset include Date, High, Low, Mid, Last, Bid, Ask, Volume from 2016 onwards.

We decided to drop this idea because of the short timeline of the currency, and a lack of other variables that can explain the prices.

1. **Dataset: NHL Playoffs**

Link: http://www.nhl.com/stats/

St. Louis Blues won the NHL Stanley cup in 2019 beating all odds, our idea was to perform analytics on the available data to understand what led to their win. NHL provides 100 years of player, team and game stats - <https://www.nhl.com/news/nhl-historical-stats-press-release/c-291394420>.

We decided to not move on with this dataset because of issues downloading the data from the website, lack of expertise with the sport in the team and high number of variables that we may not have the time to analyze.