Project Proposals

Proposal 1: - Explanatory Analysis of Traffic pullover pattern for Florida v/s Montana
Data set: https://openpolicing.stanford.edu/data/

Proposal 2: - Exploratory Analysis of top viral YouTube videos posts
Data Set: https://www.kaggle.com/datasnaek/youtube-new

Proposal 3: - Analysis of China's import and export data by WTO and comparison to the US
Data Set: https://www.wto.org/english/res_e/statis_e/trade_datasets_e.htm

Project Proposal 1: - Explanatory Analysis of Traffic pullover pattern for Florida v/s Montana

Data set: https://openpolicing.stanford.edu/data/

1. What is the problem you want to solve?

Explanatory and Exploratory Analysis of Traffic stop data and find patterns and correlation with the violation and final outcome for Florida v/s Montana.

We will explain and assess the manner and extent to which age and race play a role in traffic stops and the final outcome result. It is also interesting to explore if weather, population density plays any role in the traffic stops patterns and frequency.

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

More than 20 million Americans are stopped each year for traffic violations, making this one of the most common ways in which the public interacts with the police (Langton and Durose, 2013). This Explanatory analysis can help people understand what are the most common time, type of traffic stops and most common possible outcome. The possible outcome can help build confidence in the police and improve the law and order also.

3. What data are you going to use for this? How will you acquire this data?

Stanford is analyzing a unique dataset detailing more than 60 million state patrol stops conducted in 20 states between 2011 and 2015. This dataset is compiled through a series of public records requests filed with all 50 states, and redistributing these records in a standardized form to facilitate future analysis. The Stanford Open Policing Project data are made available under the Open Data Commons Attribution License for educational project.

Data Source and sample: https://openpolicing.stanford.edu/data/

4. In brief, outline your approach to solving this problem (knowing that this might change later)

- Data volume: - I plan to use 3 heavily populated cities of Florida (Miami, Orlando, Tampa) and compare with the entire state of Montana. I will use one year's data of a state for training, and one year's for validation and one year's data for prediction. Approx. 500K records will be used.

- Data manipulation using Pandas and other Python packages: -
Work with missing or invalid values, Data wrangling steps including filter data by year, clean messy and incomplete data, group and aggregate data

- Data analysis and visualization: - Data Storytelling using mapplotlib python package primarily focusing on the most common type of Bar chart, line chart, Scatterplot, histogram, density plot

5. What are your deliverables?

The proposal will be part of a github repository for my project. All code, data sets and further documentation I write will be added to this repository.
I will also create a Pdf project report for the entire project outcome and observations.
Project report will highlight comparison of heavy density (FL) v/s remotely populated areas (MT) including below analysis

- What is the most common cause of traffic pullover?
- What is most common outcome of traffic stop.
- Who are the repeat offenders and where are most common stops?
- What is the role of age, race and time in the traffic stops?
- What is the most probable time of traffic stop during the day /month /Year?
- How are millennials doing compared to Gen-x and is Florida really senior friendly?
- Who are most likely let go with just a warning.

Project Proposal 2: - Exploratory Analysis of YouTube videos posts
Data Set: https://www.kaggle.com/datasnaek/youtube-new

1. What is the problem you want to solve?

Explanatory and Exploratory Analysis of trending YouTube videos posted to YouTube from the US based users.

Youtube is a huge pool of data.
The total number of people who use YouTube – 1.3+ B.
300 hours of video are uploaded to YouTube every hour!
Over 800 million unique users visit YouTube each month.
Over 3.5 billion hours of video are watched each month on YouTube.

Millions of YouTube videos are added every month. Many of them start trending and a very few become viral quickly. Apart from great content what are the other facts that make a video viral?

We will explore and assess how people interact with a video once they watch it.
It is also interesting to explore if there is any pattern to predict a viral video early?

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

Worldwide, YouTube is the biggest and most used video platform. Many of the Media channels, Marketers and celebrity use YouTube as a prime source of distribution and marketing.

This Explanatory analysis can help them understand what are the of the most common things in viral videos. by Categorizing YouTube videos and Analyzing what factors affect how popular a YouTube video will be, the possible outcome can help user and marketers optimize the details and post videos more efficiently.

3. What data are you going to use for this? How will you acquire this data?

I will utilize the kaggle data set collected from the Top (up to) 200 listed trending YouTube videos every day in the United States.

Data Source and sample: https://www.kaggle.com/datasnaek/youtube-new

4. In brief, outline your approach to solving this problem (knowing that this might change later)

- Dataset: - I plan to use the US data only for this exercise.

- Data manipulation using Pandas and other Python packages: -

  Work with missing or invalid values, Data wrangling steps including filter data by year, clean messy and incomplete data, group and aggregate data

- Data analysis and visualization: - Data Storytelling using mapplotlib python package primarily focusing on the most common type of Bar chart, line chart, Scatterplot, histogram, density plot

5. What are your deliverables?

The proposal will be part of a github repository for my project. All code, data sets and further documentation I write will be added to this repository.
I will also create a Pdf project report for the entire project outcome and observations.
Project report will highlight trends of different factor in a viral video including below analysis

- How long it takes for a video to start tending after publishing
- Top Category and title of viral videos
- Average views per comment, per like, per dislike
- Most common tag and Avg/total views by tag
- What is most common Title length in words
- What is the most probable time of traffic stop during the day /month /Year?
- Comparing No of Views v/s Likes v/s comment
- Most common word used in title/description

Project Proposal 3: - Explanatory Analysis of US and China import and export data by WTO
Data Set: https://www.wto.org/english/res_e/statis_e/trade_datasets_e.htm

1. What is the problem you want to solve?

Explanatory and Exploratory Analysis of import and export business of US v/s China

US and China are 2 top economies of the world but with entirely different areas of business dominance. This analysis will determine what are driving factors for growth and which business segments are critical to growth for each country. What are the key strengths segment and what are the areas with scope for improvement?

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

This study analysis is for common people to demonstrate them how data science is used in demonstrating the key economic figure. This Explanatory analysis can help them understand what are the common pattern in us and China import and export business and how is the US performing in recent past.

3. What data are you going to use for this? How will you acquire this data?

I will utilize the free WTO dataset provided from the year 2005 till the year 2016 for import and export trade revenue.

Data Source and sample: https://www.wto.org/english/res_e/statis_e/trade_datasets_e.htm

4. In brief, outline your approach to solving this problem (knowing that this might change later)

- Dataset: - I plan to use only US and China data for this exercise.

- Data manipulation using Pandas and other Python packages: -
Work with missing and invalid values, Data wrangling steps including filter data by year, clean messy and incomplete data, group and aggregate data based on logical categories

- Data analysis and visualization: - Data Storytelling using mapplotlib python package primarily focusing on the most common type of Bar chart, line chart, Scatterplot, histogram, density plot

5. What are your deliverables?

The proposal will be part of a github repository for my project. All code, data sets and further documentation I write will be added to this repository.
I will also create a Pdf project report for the entire project outcome and observations.
Project report will highlight trends of different factor in US and China trade data including below analysis

- What is the trend of net trade deficit of US compared to China?
- How is export growth compared to import growth over the years.
- How is US-China bilateral trade doing in recent past?
- How are top US trade partners doing business with China?
- Are US and China business competitor or trade partners?
- Which trade segment or partners the US should focus to add another 1% to the growth for 2018.
