**Project Proposal**

**Project title:**

**Bike Sharing System**

* A bicycle-sharing system, bike share program, public bicycle scheme, or public bike share (PBS) scheme, is a shared transport service in which bicycles are made available for shared use to individuals on a short-term basis for a price or free.
* Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.
* Goals of Bike sharing system:
* People use bike-share for various reasons. Some who would otherwise use their own bicycle have concerns about [theft](https://en.wikipedia.org/wiki/Theft) or [vandalism](https://en.wikipedia.org/wiki/Vandalism), [parking](https://en.wikipedia.org/wiki/Parking) or storage, and maintenance.
* Most large-scale urban bike sharing programmes have numerous bike check-out stations, and operate much like [public transit](https://en.wikipedia.org/wiki/Public_transport) systems, catering to tourists and visitors as well as local residents. Their central concept is to provide free or affordable access to [bicycles](https://en.wikipedia.org/wiki/Bicycle) for short-distance trips in an [urban area](https://en.wikipedia.org/wiki/Urban_area) as an alternative to private [vehicles](https://en.wikipedia.org/wiki/Vehicle), thereby reducing [congestion](https://en.wikipedia.org/wiki/Traffic_congestion), [noise](https://en.wikipedia.org/wiki/Noise), and [air pollution](https://en.wikipedia.org/wiki/Air_pollution).
* Bicycle-sharing systems have also been cited as a way to solve the "[last mile](https://en.wikipedia.org/wiki/Last_mile_(transportation))" problem of public transit networks. However, dock systems, serving only stations, resemble [public transit](https://en.wikipedia.org/wiki/Public_transit), and have therefore been criticized as less convenient than a privately owned bicycle used door-to-door.

**Problem Statement and business use case**

* BoomBikes, a US bike sharing system faced considerable dip in the business when worldwide pandemic which has now made it difficult for the business to operate.
* Now in order to tackle this, the company has decided to form a strategy which will be able to provide better and innovative services to their customers along with keeping a check on the acceleration of their revenue generation.

**Data Science Workflow**

1. **Measures as recorded**

* instant: record index
* - dteday: date
* - season: season (1: springer, 2: summer, 3: fall, 4: winter)
* - yr : year (0: 2011, 1:2012)
* - mnth : month ( 1 to 12)
* - hr: hour (0 to 23)
* - holiday: weather day is holiday or not
* - week day: day of the week
* - working day: if day is neither weekend nor holiday is 1, otherwise is 0.
* - weathersit :
  + 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - * + -2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  + 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain +Scattered clouds
  + 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
* - temp: Normalized temperature in Celsius. The values are divided to 41(max)
* - atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
* - hum: Normalized humidity. The values are divided to 100 (max)
* - windspeed: Normalized wind speed. The values are divided to 67 (max)
* - casual: count of casual users
* - registered: count of registered users
* - cnt: count of total rental bikes including both casual and registered
  + - * + And we want to predict the count of bikes

1. **Basic assumptions.**

* Data shared with us has been collected over a span of 2 years ( i.e. 2018-2019) across four different seasons.
* We want to build a multiple linear regression model to predict the demand of bikes weighing in the various factors.
* We want to analyze the different factors and understand which factor would be impacting bike demands.

1. **Solutions:**

* Identify which factors are important in predicting bike demand in US. Using correlation analysis and EDA.
* To build a model to predict demand in future based on these attributes.

1. **Alternative solutions, is to work on some exploratory data analysis.**

* At what time and month in a year does the bike demand spikes?
* At what temperatures people prefer to use the bike service?
* In which season does the demand usually goes up?
* On which day of the week is the bike demand usually high?
* Does the bike demand vary much on nationwide holidays in a year as compared to nonholidays?
* Does the bike demand vary on weekends as compared to weekdays?
* How many bikes a used on casual basis?
* How many bikes are used on a registered basis?
* In which year was the bike demand more?

1. **Data Analytics**

* Descriptive analysis – count of bikes utilized are affected by the varying weekdays and holidays, temperatures and usage patterns as registered and casual.
* Diagnostic analysis -Registered users may find bike sharing facility as the means to reach their work location or a daily necessary commute.
  + During weekends and holidays, the utilization of bike service is more, which points to the direction in which people tend to enjoy riding these bikes as a recreation activity which is also healthy.
  + In lower temperatures, the service is not utilized much due to the fact that is it too cold and too uncomfortable to ride bikes in heavy clothing.
* Predictive analysis- effort to increase casual users by various means will prove beneficial for the company.
* Prescriptive analysis-
* By proving daily and group offers, better tracking system, easier accessibility, etc will lead to increase in the number of casual uses along with the registered ones as well.
* Campaigning in order to make people aware of the health benefits this service brings along with it will increase customer base thus more revenue.
* Try to sanitise vehicles after every use to show your 2 pennies in making it safe in this ongoing pandemic.

1. **What can we do to fix it?**
   * Awareness campaigns.
   * More offers.
   * Defining target market.
   * Better service than competitors.
   * More junctions closer to public areas and office hubs.
   * Sanitization
   * 24/7 online support
   * Clarity on use for casual users

**Data Collection**

* Data set is in a csv form.
* Data set consists of data of span of 2 years (2018 & 2019)
* Data shape = 730 rows and 16 columns
* It has size of 56.1Kb
* Data columns are: instant', 'dteday', 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit', 'temp', 'atemp', 'hum', 'windspeed’, ‘casual', 'registered', 'cnt'.

**Data preprocessing preparation and Feature Engineering.**

* Identifying data quality.
* Check for data inconsistency, data types and no encoding issues.
* Handle missing values.
* Treating categorical values
* Treating outliers
* Data scaling
* Feature engineering.

**Machine Learning and model selection**

* Multiple liner regression- base line
* Ridge regression, lasso regression, stochastic linear regression.

**Model Persistence and deployment.**

* Building and API to expose the model to predict future inputs.
* Deployment on local machine- using Ngrok and flask
* Cloud AWS, GCP.