**Pittsburgh bike share Demand Prediction**

**Introduction:**

Bike sharing systems have been increasing in demand over the past two decades as a result of rapid advancements in technology. Healthy Ride is a public bicycle sharing system that serves parts of Pittsburgh to fulfill the growing need for changes in mobility pattern. With 100 stations, the city has plans to add 75 more stations in 2019, including several electric bikes to help riders navigate Pittsburgh’s hilly geography, located throughout the city.

**Problem:**

Bike-sharing systems are used world-wide. Given that the system tends to be unbalanced, particularly

at peak demand hours, there are challenging analytical issues such as accurately predicting the demand.This project explores on predicting the total count of bikes rented during each hour covered by the Healthy Ride operated by Pittsburgh Bike Share.

**Clients**:

Bike sharing operators can use this model to proactively shape the mobility market by forecasting demand prediction and  to meet customer expectations.

**Datasets and Approach:**

**Healthy Ride operated by Pittsburgh Bike Share data:**

Healthy Ride is the latest mode of transit in Pittsburgh, PA. The  combined dataset has roughly about 285455 rows of data. I have selected csv files from 2015, Quarter 2 to 2018, Quarter 4.

<https://data.wprdc.org/dataset/healthyride-trip-data>

This dataset includes these columns:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Trip id | Starttime | Stop  time | Bikeid | Trip  duration | From station id | From station name | To station id | To station name | User  Type |

Using station Id as the join clause, Longitude and Latitude can be merged from the bike station dataset as below:

<https://healthyridepgh.com/data/>

This dataset includes these columns:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| StationNum | StationName | RackQnty | Latitude | Longitude |

**Distance data:**

Distances between stations are not included in Healthy Ride Bike share's data release. To consider the  bicycling distances between pairs of lat/lon coordinates, we can use Google Maps [distances API](https://developers.google.com/maps/documentation/distance-matrix/intro)

According to the [Usage and billing](https://developers.google.com/maps/documentation/distance-matrix/usage-and-billing) fine print, it costs about 5.00 USD per 1000 requests. With 50 stations in Pittsburgh, It might cost about $10 to $15  for total requests. Else, I could also use [Haversine’s](https://andrew.hedges.name/experiments/haversine/)  formula  to calculate the distance as explained below:

<https://stackoverflow.com/questions/19412462/getting-distance-between-two-points-based-on-latitude-longitude>

Here's what the first 5x5 section of the distance matrix might look like:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Station1** | **St-02** | **st-3** | **st4** |  |
| st -01 | 0 | ~14 | ~167 | ~156 |  |
| **sT-02** | ~12 | 0 | ~67 | ~867 |  |
| **sT-03** | ~422 | ~23 | 0 | ~445 |  |

Given these distances, we can also compute Speed since we already know the time taken.  To know how the number of rides changes with temperature and precipitation, we can take a look at weather and Storm data as below:

**Weather data:**

Url**:** [**https://www.ncdc.noaa.gov/cdo-web/search**](https://www.ncdc.noaa.gov/cdo-web/search)

The columns included in this dataset are:

* Station Name,
* WSF2 - Fastest 2-minute wind speed
* WSF5 - Fastest 5-second wind speed
* WT03 - Thunder
* WT04 - Ice pellets, sleet, snow pellets, or small hail"
* PRCP - Precipitation
* WT08 - Smoke or haze
* WT09 - Blowing or drifting snow
* WDF2 - Direction of fastest 2-minute wind
* AWND - Average wind speed
* WDF5 - Direction of fastest 5-second wind
* PGTM - Peak gust time
* WT01 - Fog, ice fog, or freezing fog (may include heavy fog)
* TMAX - Maximum temperature
* WT02 - Heavy fog or heaving freezing fog (not always distinguished from fog)
* TAVG - Average Temperature.
* TMIN - Minimum temperature

**Storm DATA**:

Url link: [www.ncdc.noaa.gov/stormevents/](http://www.ncdc.noaa.gov/stormevents/)

The columns included in this dataset are:

     ['EVENT\_ID', 'CZ\_NAME\_STR', 'BEGIN\_LOCATION', 'BEGIN\_DATE', 'BEGIN\_TIME',  'EVENT\_TYPE', 'MAGNITUDE', 'TOR\_F\_SCALE', 'DEATHS\_DIRECT', 'INJURIES\_DIRECT', 'DAMAGE\_PROPERTY\_NUM', 'DAMAGE\_CROPS\_NUM',   'STATE\_ABBR', 'CZ\_TIMEZONE', 'MAGNITUDE\_TYPE', 'EPISODE\_ID', 'CZ\_TYPE', 'CZ\_FIPS', 'WFO', 'INJURIES\_INDIRECT', 'DEATHS\_INDIRECT', 'SOURCE',  'FLOOD\_CAUSE', 'TOR\_LENGTH', 'TOR\_WIDTH', 'BEGIN\_RANGE', 'BEGIN\_AZIMUTH', 'END\_RANGE', 'END\_AZIMUTH', 'END\_LOCATION', 'BEGIN\_LAT', 'BEGIN\_LON', 'END\_LAT', 'END\_LON', 'EVENT\_NARRATIVE',  'EPISODE\_NARRATIVE', ABSOLUTE\_ROWNUMBER']

 353 events were reported between 04/01/2015 and 12/31/2018 (1371 days). Event types   with the number of occurrences are below:

* Thunderstorm Wind   124
* Flash Flood          119
* Hail                  42
* Flood                 35
* Heavy Rain            14
* Winter Weather         6
* Strong Wind            3
* Winter Storm           2
* Heavy Snow             2
* Cold/Wind Chill        2
* Ice Storm              1
* Lightning              1
* Debris Flow            1
* Tornado                1

**Bike score data:**

I will also add bike score data  from [https://www.walkscore.com](https://www.walkscore.com/).  and transit score from <https://www.walkscore.com/professional/public-transit-api.php>

Bike Score service measures whether a location is good for biking on a scale from 0 - 100 based on four equally weighted components:

* Bike lanes
* Hills
* Destinations and road connectivity
* Bike commuting mode share

|  |  |
| --- | --- |
| **Bike Score** | **Description** |
| **90–100** | **Biker's Paradise**  Daily errands can be accomplished on a bike. |
| **70–89** | **Very Bikeable**  Biking is convenient for most trips. |
| **50–69** | **Bikeable**  Some bike infrastructure. |
| **0–49** | **Somewhat Bikeable**  Minimal bike infrastructure. |

**Transit Score :**

The score API call returns a Transit Score for a location that is an integer between 0 and 100.Transit Score is a patented measure of how well a location is served by public transit on a scale from 0 to 100.

|  |  |
| --- | --- |
| **Transit Score** | **Description** |
| **90–100** | **Rider's Paradise**  World-class public transportation. |
| **70–89** | **Excellent Transit**  Transit is convenient for most trips. |
| **50–69** | **Good Transit**  Many nearby public transportation options. |
| **25–49** | **Some Transit**  A few nearby public transportation options. |
| **0–24** | **Minimal Transit**  It is possible to get on a bus. |

**Data Fields/Features that can be extracted by combining datasets:**:

* datetime - hourly date + timestamp
* Distance between start & stop stations
* Speed
* Bike Score
* Transit Score
* season -  1 = spring, 2 = summer, 3 = fall, 4 = winter
* holiday - whether the day is considered a holiday
* working day - whether the day is neither a weekend nor a holiday
* weather - 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: All the event types listed in the Storm dataset
* temp - temperature in Celsius
* humidity - relative humidity
* windspeed - Avg wind speed
* casual - number of non-registered user rentals initiated
* registered - number of registered user rentals initiated
* count - number of total rentals

I will segregate the data into train set(1st 20 days of any month) and test set(20th onwards of any given month).I would be applying the regression techniques to predict the exact number of bikes that will be rented.

**Deliverables:**

Powerpoint presentation, website and Github.