**Project 1 Report:**

**Exploratory Data Analysis**

Allen Martillano

martillanoallen@gmail.com

**Abstract –** In this report, the purpose of this data analytics project is to perfrom descriptive statistics with exploratory data analysis using tools like excel. Futhermore, using data from 2011-2012 of the bike sharing dataset.

**Introduction**

First, we must uncover the dataset and it’s task description. Then we must discover the descriptive statistics on If they are numerical, or categorical variables. Then we must create a frequency table on considering one categorical variable, and then creating a pivot table. For the numerical variables we must uncover their statistics (min,max,avg,etc). Then create a histogram, and a box plot for one numerical variable. After we must perform simple liner regression, and multiple linear regression models. To determine which model best fits to be able to predict the number of bike rentals (hourly and daily) based on previous data on enviornmental conditions and other variables. Further analysis on the excel file.

Link to the dataset: [UCI Machine Learning Repository: Bike Sharing Dataset Data Set](https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset)

**Exploratory Data Analysis**

*Dataset and task description – describe the data analytics task for the chosen dataset*

The task associated with this dataset is Regression because the values in the target data (numerical data) are continuous (minimum to maximum).

We can predict continuous variables like the prediction of bike rentals. For instance, we can forecast these results from past data, and we can use the regression algorithm. Once complete, we can predict bike rentals for future days. This algorithm can be divided into linear and non-linear regression. Although, various factors change through continuous data. The corresponding weather, seasonal information, and travel are all input values that correspond with the continuous output of the number of bike rentals.

*How is the dataset used to solve a business problem?*

We can predict the number of bike rentals (hourly and daily) based on previous data on environmental conditions and other variables.

*Show the dataset characteristics, namely the number of instances, the number of numerical and categorical variables, and the description and the domain for each.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hour dataset** | | | **Day dataset** | | |  |
| *number of instances* | *number of numerical variables* | *number of categorical variables* | *number of instances* | *number of numerical variables* | *number of categorical variables* |  |
| 17379 | 8 | 7 | 731 | 7 | 7 |  |

|  |  |
| --- | --- |
| **Hour dataset numerical viarables** | |
| *Description* | *Domain* |
| dteday : date | date |
| season : season (1:winter, 2:spring, 3:summer, 4:fall) | categorical |
| yr : year (0: 2011, 1:2012) | categorical |
| mnth : month ( 1 to 12) | categorical |
| holiday : weather day is holiday or not | categorical |
| weekday : day of the week | categorical |
| workingday : if day is neither weekend nor holiday is 1, otherwise is 0. | categorical |
| 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 | categorical |
| temp : Normalized temperature in Celsius. The values are divided to 41 (max) | numerical |
| atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max) | numerical |
| hum: Normalized humidity. The values are divided to 100 (max) | numerical |
| windspeed: Normalized wind speed. The values are divided to 67 (max) | numerical |
| casual: count of casual users | numerical |
| registered: count of registered users | numerical |
| cnt: count of total rental bikes including both casual and registered | numerical |

Further analysis on the excel file.