Chart, histogram

Description automatically generated

Stats: Casual

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

* Response Variable distribution
* Predictors distribution – pairplot

Diagram

Description automatically generated

Diagram, schematic

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Introduction**

Public bike sharing has experienced a sharp increase on a global scale as an ingenious mobility solution. Although bike-sharing systems offer quick, affordable, and environmentally beneficial transportation, their unique features have negative effects on both riders and operators. In contrast to traditional public transit (such as buses and subways), which adheres to a set timetable and predetermined routes, bike-sharing offers an on-demand transportation. This leads to an uneven distribution of bikes brought on by fluctuating demand and supply.

Effective bike rebalancing solutions are required to address this bicycle imbalance issue, which heavily relies on bicycle mobility modelling and prediction. Due to the imbalance of bicycles, bikeshare towns must use expensive redistribution of bikes, which is normally carried out by trucks or trailers travelling throughout the city and relocating bikes between stations. Studies have been done to improve these bike redistribution procedures based on bicycle mobility models and predictions in order to maximise service availability and decrease redistribution cost.

**Data Set**

The primary data set comes from the Capital Bikeshare system in Washington, D.C., USA, and is based on a two-year historical log for the years 2011 and 2012 that is publicly accessible at http://capitalbikeshare.com/system-data. We have used the aggregated at the daily level available at <https://www.kaggle.com/datasets/marklvl/bike-sharing-dataset>.

The dataset contains per day casual user count with 731 entries, processing 14 attributes – 13 independent and one dependent form the part of our Regression Analysis. The dataset contains season, year, month, hour, whether or not it was a holiday, weekday, working day, the weather conditions, temperature, feeling temperature casual variable count and date information. Date does not provide relevant information to generate a model to predict the number of casual users.

The main goal was to create a superior statistical model to forecast the number of casual users that would rent a bike on a particular day and to comprehend the patterns and factors that affected the number of rented bicycles on a given day.

Our goal is to predict the number of casual users, **casual,** using a suitable linear regression model.

**Literature Review**

Since the establishment of the first bike sharing system in the Netherlands in the 1960s, there have been four generations of bike sharing (DeMaio, 2009) (Shaheen, 2010). Since the release of the third generation, bike sharing has grown in popularity. The automatic transaction kiosk at each station and identifiable bike sharing users can be used to describe the third generation of bike sharing. Around the world, these methods have achieved a fair amount of success. Fourth-generation bike sharing programmes featuring improved docking stations, bike redistribution, interaction with other means of transportation, and electrical bikes have been built in Copenhagen and Madrid (DeMaio, 2009) (Shaheen, 2010).

Numerous studies have recently employed conventional surveys to ascertain the elements that would encourage urban communities to adopt bike sharing (Bikeshare, C., 2013) (Share, A. B., 2011). An invaluable resource for learning more about how bike sharing is used in the city is the automatic data collected from docking stations. Numerous studies have identified factors that affect the use of bike sharing and have attempted to forecast bike sharing flow using various urban factors, including: population, jobs, bicycle lanes, proximity to public transportation, density of bike sharing stations, altitude, retail shops, etc. (Faghih-Imani A. E.-G., 2014) (Rixey, 2013) (Wang, G., & JE, 2012). These studies' use of daily, monthly, or annual aggregated data can obscure the variation of everyday bike sharing usage (Rixey, 2013) (Wang, G., & JE, 2012). In Barcelona and Seville, Spain, Hampshire used sub-city district level aggregated hourly arrival and departure rates to study the built environment and bike sharing utilisation (Faghih-Imani A. H., 2017). They discovered that the density of bike sharing stations, the capacity of the stations, and the number of sites of attraction are crucial variables in explaining the arrival and departure rates of bike sharing. But rather than using bike sharing flows at the station level, their study aggregated the flows at the level of sub-city districts, which was less meaningful.

**Research Problem**

In this project, our goal is to draw inferences from the dataset on bike sharing in order to maximise its worth and obtain information that will help the system function more effectively given the set of factors.

We try to answer the following questions.

* Is temperature significant predictor for casual users?
* Does season influence number of casual users? If yes, which two seasons show significant difference in average casual users?
* What is the effect of season after adjusting for temperature?

Our goal for this project is two-fold

* To answer the above questions give justification for our answers
* Yo understand the patterns and variables that affected the number of leased bicycles on a given day and to develop a superior statistical model to predict the number of casual customers who will rent a bike on a specific day

**Examining the Data**

**Description of Data**

**Name of Dataset:** Bike Sharing in Washington D.C. Dataset

**Number of Observations:** 731

**Attribute Information:**

* **instant**: Record index
* **dteday**: Date
* **season**: Season (1:springe, 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 (extracted from [Holiday Schedule](http://dchr.dc.gov/page/holiday-schedule))
* **weekday**: Day of the week
* **workingday**: If day is neither weekend nor holiday is 1, otherwise is 0.
* **weathersit**: (extracted from [Freemeteo](http://www.freemeteo.com/))
  + 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 derived via (t-t*min)/(t*max-t*min), t*min=-8, t\_max=+39 (only in hourly scale)
* **atemp**: Normalized feeling temperature in Celsius. The values are derived via (t-t*min)/(t*max-t*min), t*min=-16, t\_max=+50 (only in hourly scale)
* **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

**Target:** casual

**Preliminary Analysis**

**Descriptive Statistics**

R Studio was used for the pre-processing of our rental cycle dataset. The read.csv() function was used to load the CSV file.

The R is.na() function is used to examine the missing data. The results showed that our dataset contained no missing values.

Descriptive Statistic measures including the mean, variance, standard deviation, median, and maximum and minimum values were calculated for the target variable, casual, as well as the continuous predictors: temp, atemp, hum and windspeed. The result is shown below.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

We then checked the distribution of our target variable casual using a histogram and a box plot. Both the plots are shown below.

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

As a next step to our initial analysis, we visualized our categorical features using scatter plots and box plots. These plots are shown below.

Diagram, schematic

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Define Research problem

* Questions
  + Is temperature significant predictor for casual users?
  + Does season influence number of casual users? If yes, which two seasons show significant difference in average casual users?
  + What is the effect of season after adjusting for temperature?
  + Domain understanding gives – temperature, humidity, windspeed and
* Goal:
  + To answer above questions
  + To build a model to predict casual
* Analysis
  + Data Exploration (details like what kind of data we have)
  + Data Transformation
  + EDA (inference from data)
    - Categorical Variable analysis
      * Boxplot & Scatter plot
    - Continuous variable analysis
      * Hist plot
      * Corr
      * Pair

- Modelling

* Hypothesis testing – CI, Prediction Interval
* Check for model assumptions
* Conclusion
* Variable selection (Stepwise)
* Model Assumption
* Influential points
* VIF
* Model Final – CV
* Random Forest

# Bibliography

Bikeshare, C. (2013). *Capital bikeshare member survey report.* Washington, DC.

DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of public transportation*, 3.

Faghih-Imani, A. E.-G. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of transport geography*, 306-314.

Faghih-Imani, A. H. (2017). An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville. *Transportation Research Part A: Policy and Practice*, (pp. 177-191).

Rixey, R. A. (2013). *Station-level forecasting of bikesharing ridership: Station network effects in three US systems.* Transportation research record.

Shaheen, S. A. (2010). *Bikesharing in Europe, the Americas, and Asia: past, present, and future.* Transportation research record.

Share, A. B. (2011). *Melbourne bike share survey.* Melbourne: Melbourne: Alta Bike Share.

Wang, X. (., G., S., & JE, H. A. (2012). *Modelling bike share station activity: the effects of nearby businesses and jobs on trips to and from stations.* TRB’s 92nd Annual Meeting and Publication in the Transportation Research Record: proceedings.