**Final Report**

**GROUP 2 MID-TERM PROJECT:**

**Linear Model LM - Predicting Bike Sharing Demand in the Cities**

**Author: Modupeola Fagbenro, Crystal Kao, Guruksha Gurnani, Shanun Randev,**

**Introduction**

**Background of the study**

The given data set encompasses information about bike sharing, a rapidly growing sustainable mode of transportation. Bike-sharing systems have been gaining more popularity in urban areas than in rural ones due to the convenience and ease of traveling. This eco-friendly system allows individuals to rent bikes for short periods typically on an hourly basis, making them an efficient choice of commute, and leisure and reducing environmental impact.

Our data spans over 10,000 rows across 15 columns– season, Holiday, working day, Weather, Temperature, Atemp, Humidity, windspeed, casual, registered, count, datetime.

As a direct quote from the data source:

“ Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city…” (Cukierski, 2014)

The data is specifically collected from the Capital Bikeshare program in Washington, D.C for 2011-2012.

**Problem Statement**:

In bustling urban centers characterized by heavy traffic and relatively short distances between points of interest, bike sharing has emerged as a viable alternative mode of transportation. Bike sharing allows individuals to conveniently rent bicycles from designated stations, without the requirement of returning them to the same station. Instead, there are numerous drop-off locations scattered throughout the city, offering increased flexibility in returning bicycles.

**Research Aim and Objectives**

The objective of this research is to check factors that can influence Bike Sharing:

1. **Checking Number of Urban Rentals** Higher casual or registered population density in urban areas typically makes bike sharing more appealing due to shorter distances between destinations and increased demand for alternative transportation options.
2. **Weather Conditions:** Weather, such as temperature, humidity, or extreme heat, can impact bike sharing. Cities with more favorable weather conditions may see higher bike-sharing usage.
3. **High-year Rentals:** Theincrease in bikes year in and year out can be a result of good weather conditions or an increase in awareness of alternative means of transportation such as bike sharing.

**Research SMART Questions**:

**Specific Measurable Achievable Relevant and Time-Bound Questions (SMART):**

Here are the SMART Questions, our analysis aims to answer:

1. Which variables are useful for predicting the number of bikes used in an hour?
2. How does temperature affect the number of bikes in use?
3. When do casual bikers increase the most?

Applying the SMART framework to our questions allows us to set clear objectives, measure progress, and ensure that the analysis provides meaningful insight for decision-makers/stakeholders in the context of bike sharing.

**Dataset Fields** :

* Datetime – The hour, date, and timestamp of each observation.
* Season: Categorical variable indicating the season, which can affect bike usage (e.g., 1-spring, 2-summer, 3-fall, 4-winter).
* Holiday: It’s binary variable indicating whether it's a holiday (1 for yes, 0 for no)
* Working Day: It's a binary variable indicating whether it's a working day (1 for yes, 0 for no).
* Weather: Categorical variable representing different weather conditions, influencing bike ridership (e.g., clear, cloudy, rainy).

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

* Temperature (temp): Temperature in Celsius, affecting user preferences for bike rides.
* Atemp : "Feels Like" Temperature (atemp): This column represents the "feels like" temperature, which takes into account factors such as wind chill and humidity to provide a more accurate perception of temperature. It can affect user comfort and, consequently, bike usage.
* Humidity: Humidity level, which can impact the comfort of cycling.
* Windspeed: Windspeed, another weather-related factor that affects bike riding conditions.
* Casual: The number of casual (non-registered) bike rentals.
* Registered: The number of registered bike rentals, indicating regular users of the bike sharing system.
* Count: The total count of bike rentals, which is the sum of casual and registered rentals, serving as the primary target variable for analysis and prediction.

Based on the above information, this data set allows us to explore how different factors including but not limited to weather conditions, time of the day, and humidity can impact bike-sharing trends and use these insights to improve bike-sharing systems.

**Exploratory Data Analysis and Methodology**

Performing exploratory data analysis (EDA) on a bike-sharing dataset is an essential step to gain insights and understand the data before diving into more advanced analyses. Here's a step-by-step guide to performing EDA on a bike-sharing dataset and what methods we used:

**EDA Steps Conducted:**

1. **Data Collection:** Start by obtaining the bike-sharing dataset from Kaggle. <https://www.kaggle.com/competitions/bike-sharing-demand/data>
2. **Data Loading:** Import the dataset into R studio data analysis environment.
3. **Load Useful Packages:** The packages used in this analysis include: tidyverse, psych, ggthemes, ggplot2, dplyr, tidyr, corrgram, lubridate, caret, corrplot
4. **Data Overview:** Check the first few rows of the dataset to understand its structure and the kind of information it contains. Next check the data types of each column (e.g., categorical datatypes and numerical datatypes) to ensure they are correctly interpreted. Finally, inspect the columns (variables) in the dataset and their names.
5. **Data Cleaning:** Handle missing data by either imputing missing values or removing rows/columns with missing data. This data file luckily did not include any missing values or rows.
6. **Summary Statistics:** Calculate summary statistics like mean, median, standard deviation, and percentiles for numerical variables. This provides an initial understanding of the data distribution. For categorical variables, calculate frequency counts and proportions.
7. **Data Visualization:** Create visualizations to gain insights into the data. Some plots we used include scatterplots, box plots, bar charts, and correlation matrices.
8. **Distributions and Patterns:** Examine the distribution of the target variable(s) (e.g., bike rentals) and potential predictor variables. For this analysis, we looked at bikers by temperature, bikers by humidity, and bikers by season, which will all be shown in the results section as they pertain to our SMART questions.

**Changes Made After EDA**

After doing some EDA, some of our questions changed to adhere more to what made more sense for our data to answer and focus in on answering more pointed questions. For example, one of our original questions was which variable has the highest effect on the count of the number of bikes used in a day? However, due to the way the data is collected and the purpose of the dataset, it made more sense to change the question to which variables are useful for predicting the number of bikes used in an hour? Since the data is collected hourly and creating a model instead of looking for a specific variable made more sense. The model can be used to predict future bike sharing rates.

**Model Selection and Hypothesis Testing**

Since we were interested in learning about which variables influence the number of bikers in a given hour, we wanted to create a linear regression model. Before doing so, we needed to format our data appropriately. We created new fields based on the ‘Datetime’ column to split out the year, day, month, hour and day of the week. ‘Season’ and ‘weather’ are categorical variables and were represented using dummy variables.

We first started out by running a linear model on the data with ‘count’ as the dependent variable and all other available variables as the predictors (this is considered Model 1). However, we excluded ‘registered’ and ‘casual’ because they sum up to the ‘count’. Based on the results of the residuals vs fitted plot, the variance of residuals increased as the values increased. A log transformation was considered to see if there was a model with more constant variance. Because of this, we also ran a version of the model with a log transformation on the ‘count’ field to see if the normality assumption would be more reasonable on the log scale (this is considered Model 2). With this new model, the residuals vs fitted plot no longer increased as the values increased, and the qq-plot also shows that normality is a reasonable assumption. Noticing that ‘holiday’ and ‘day’ were not statistically significant in this model based on the t-tests, we created another model based on the log transformed model where ‘holiday’ and ‘day’ were taken out (this model is considered Model 3). An ANOVA was then run on this model with the following hypotheses:

Null Hypothesis: Model 3 is a reasonable fit

Alternative Hypothesis: Model 2 is a reasonable fit

The p-value is greater than 0.05, so we do not reject the null hypothesis, which means Model 3 is adequate to model the rentals. There's no evidence that including ‘holiday’ and ‘day’ improves the model's fit.

An ANOVA was also run on Model 3 to confirm that all the variables included have statistical significance even after adjustment for the other variables. If any of the variables were not statistically significant, we would have continued our model selection by removing those variables. However, all the variables were statistically significant, so we continued our analysis with Model 3.

**Results**

After going through all the processes in our pipeline we are able to answer our 3 SMART questions. In the first question we need to extract the features which are useful for predicting the number of bikes used in an hourly granularity. After doing model selection based on the linear regression models, we came to the conclusion that the variables ‘holiday’ and ‘day’ are not statistically significant variables, unlike the other variables. Model 3 proved to be the most adequate model for the rental of bikes per hour due to the ANOVA we ran to compare Model 3 and Model 2. The additional ANOVA ran on Model 3 also proved that removing any more of the variables would reduce model fit. This means that all the variables other than ‘holiday’ and ‘day’ are significant and important for predicting the number of bike rentals in an hour. The final list of variables that are useful for predicting the number of bikes used in an hour are ‘season’, ‘working day’, ‘weather’, ‘temp’, ‘atemp’, ‘humidity’, ‘windspeed’, ‘year’, ‘month’, ‘hour’, and ‘dayofweek’.

In the second smart question we need to check the relationship between temperature (independent variable) and the rental counts. According to our EDA plot, the bike count increases with higher temperature but drops off once temperatures crossed 30 degrees Celsius. This makes sense considering 20-30 degrees Celsius is considered the most comfortable temperature range to be outside. The number of rental counts also increased with increasing humidity. Our final model (Model 3) also shows a slight increase in the number of rental counts when temperature increases due to the positive coefficient in the table.

In the third smart question we need to answer when do the casual (or non-registered) bikers increase the most. To answer this question, we plotted a scatterplot between temperature and the number of casual bike users. The number of casual bike rentals increased when the temperature increased but dropped off once the temperature started going beyond 30 degrees Celsius. Most of them rented bikes when the temperature was between 20-30 degrees Celsius. In Fahrenheit that is between 68-86 degrees. Again, this makes sense as people tend to go outside only when the temperature is nice and not too hot or cold.

**Conclusions/Next Steps**

In conclusion, through the process of this analysis, we were able to explore the bike sharing data provided and practiced EDA while also learning how to create models and to help answer our SMART questions. As a potential next step for future consideration, we will use the model we selected to generate predictions on a test dataset by predicting the number of bike rentals per hour. We will also try to use some more advanced techniques like hyperparameter tuning and model pruning if we get high error scores.

**References**

Cukierski, W. (2014). Bike Sharing Demand. Kaggle. https://www.kaggle.com/competitions/bike-sharing-demand/overview