



Contents

Business Understanding 4

Data Understanding 4

[Data Preparation 6](#_TOC_250011)

[Statistical Techniques for Data Analytics 11](#_TOC_250009)

[Machine Learning 14](#_TOC_250008)

[Deployment 16](#_TOC_250006)

[Conclusion 16](#_TOC_250005)

[References 17](#_TOC_250000)

[GitHub Link 17](#_TOC_250000)

**1. Business Understanding**

**Objective**

The focus of this analysis is to provide ready to work insights to this global bike rental company operating across major cities (E.G.: New York, London, Berlin, Chicago, San Francisco, and Dublin).

The three offered types of bikes—Standard, Mountain, and Electric—perfect to its diverse customer base. Now, understanding customer preferences, rental patterns, influencing factors, and so on is critical to optimizing operations and increasing profitability.

From the dataset provided, key objectives of this analysis are:

1. **Operational Efficiency**: To improve fleet management we will analyse usage patterns across cities, bike types, and weather conditions. Then, with this insights resources will be allocated more effectively.
2. **Customer Satisfaction**: I will identify factors that alters rider satisfaction to improve service offerings and guarantee the most positive customer experience.
3. **Pricing Strategy**: With the help of this analysis, we will be able to select the best price strategy for every situation in the market .

This project uses the CRISP-DM framework to ensure a structured and robust approach. Everything follows an order and we will follow these steps in order to achieve the goal of this document.

**Business Questions**

First of all, lets address the main questions this analysis will have the answer for:

1. What are the main factors that affects the cost of bike rentals? We will answer this by exploring the impact of demographics, ride characteristics, and weather conditions against this important metric; pricing.
2. How do Rider Satisfaction affects profitability? Does it have an impact on it at all? .
3. Effectively can machine learning models predict a future income or maybe being able to segment customers into groups? We will see if predictive modeling helps us in better decision-making and resource allocation.
4. What about flee management? Can we improve it?

Being able to align this project with our customer goals is the key aspect of this phase of the report.

**2. Data Understanding**

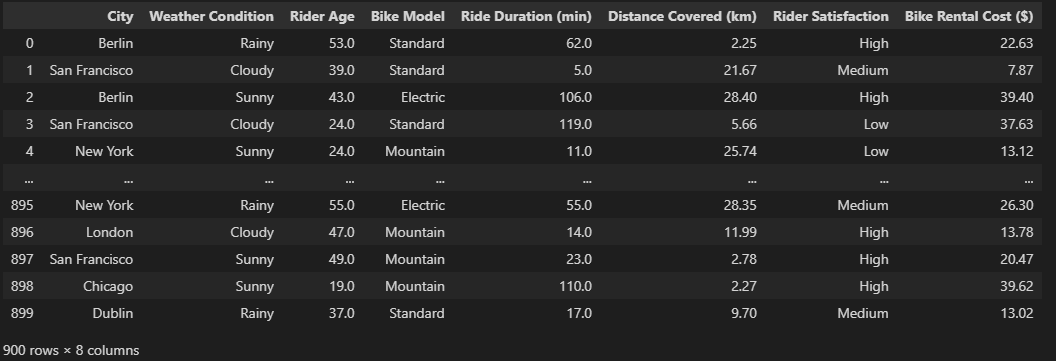
**Dataset Overview**

900 observations with attributes related to bike rentals are contained in our dataset. These attributes shows various aspects of the business (E.G.: demographics, trip details, bike models, and weather conditions). Each data point is valuable for an indispensable customer based decision-making.

Features in the dataset:

* **City**: The location of the rental, helps in identify place specific usage trends.
* **Weather**: Conditions during the ride (e.g., sunny, rainy).
* **Rider Age**: The age of the customer, this is an essential demographic feature.
* **Bike Model**: Standard, Mountain, or Electric.
* **Ride Length (minutes)**: Measures the duration of the ride.
* **Distance Covered (km)**: Total distance travelled.
* **Rider Satisfaction**: Categorical variable indicating satisfaction (E.G.: High, Medium, Low).
* **Bike Rental Cost ($)**: Price paid by the customer.

*Figure 1: dataset head.()*

**

**Initial Observations**

1. **Missing Values**:
   * Notable in **Weather**, **Rider Satisfaction**, **Rider Age**, **Ride Duration, Distance Covered**. This will require applying best practices ways of dealing with this kind of things.
2. **Outliers**:
   * I took advantage of the “SimpleImputer” to get rid of possible outliers, one example Rider Age having number below 0 or above 100, and another example could be Bike Rental Cost ($) being excessively high or low in some cases, so I just got rid of that.
3. **Data Types**:
   * Categorical and numerical features interchangeably. Categorical; **City** and **Weather.** Numerical**: Ride Length** and **Distance Covered** are numerical.
4. **Distributions**:
   * Showed not normal distributions and actually the p-value mean of the whole dataset showed a not normally distributed dataset.

**Importance of Features**

Now these are why the features have a value in answering business objective questions:

* **City** and **Weather** are crucial for understanding regional patterns and external influences. An study of these can guide fleet allocation and geographics marketing strategies.
* **Rider Age** and **Satisfaction** help profile customers, revealing preferences and things that may or not may influence satisfaction levels.
* **Distance Covered** and **Ride Length** correlate directly with pricing which could inform for future pricing strategies.
* **Bike Rental Cost ($)** this is the central variable for evaluating the financial performance. Basically what we have to look after for revenue growth.

**Preliminary stats:**

* **Average**:
  + **Rider Age**, which are the users, averages 40 years, this tells there is a balanced demographic distribution among users.
  + **Ride Duration** mean of 62 minutes, meaning moderate rentals every time.
  + **Rental Cost** 26 dollars, affordability for users.
* **Correlations**:
  + As mentioned; strong correlation between **Distance Covered**, **Bike Rental Cost and Ride Duration, showing us** that distance and time affects pricing.

With this very strong understanding of the data, the very next step is preparing data for analysis with cleaning, feature engineering, and dimensionality reduction for example.

**3. Data Preparation**

Data preparation is crucial in the CRISP-DM framework. Why? Because in this step we make the dataset clean and ready for strong analysis. Key preprocessing steps best practices for example; handling missing values, treating outliers, adjusting data types, performing exploratory data analysis, and engineering features will be done along this stage of the project.

**Data Cleaning**

1. **Missing Values**:
   * **Numerical Features**: Missing values, for example; **Rider Age, Ride Duration** and **Distance Covered** has been changed using their mean, making use of this bests practices we kept central tendency and as well as maintained dataset's demographic.
   * **Categorical Features**: Missing values in **Weather, City, Bike Model** and **Rider Satisfaction** were imputed using the mode. With this approach we minimize distortions in the categorical distributions.

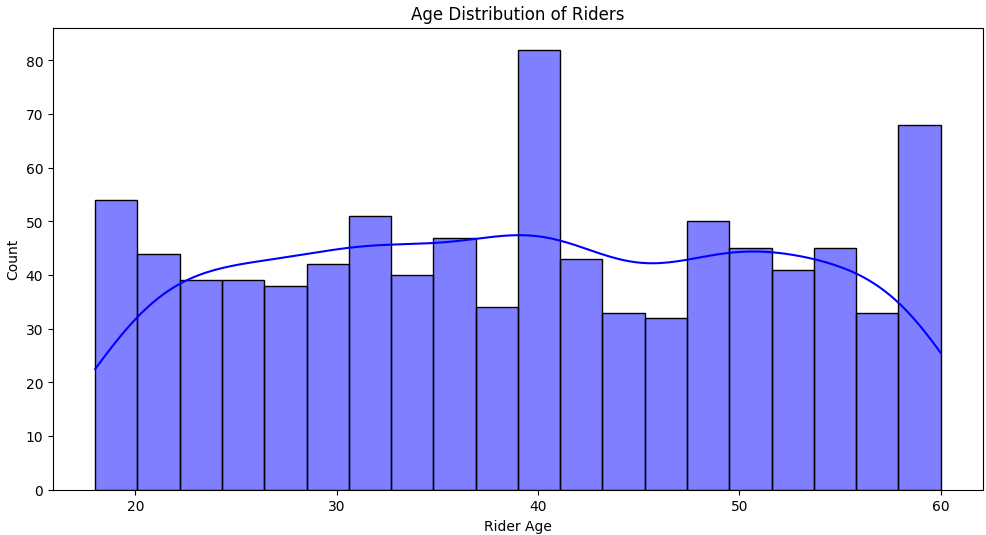
These imputation techniques were chosen for their simplicity and effectiveness, especially in this kind of dataset, a dataset with relatively small missing data handling.

1. **Data Type**:
   * **Categorical**: For best practices **City**, **Weather**, **Bike Model** and **so on** has been change to categorical.
   * **Standardization**: For later on for machine learning models. **Ride Length, Bike Rental, Rider Age** and **Distance Covered** has been standardized.

**Exploratory Data Analysis**

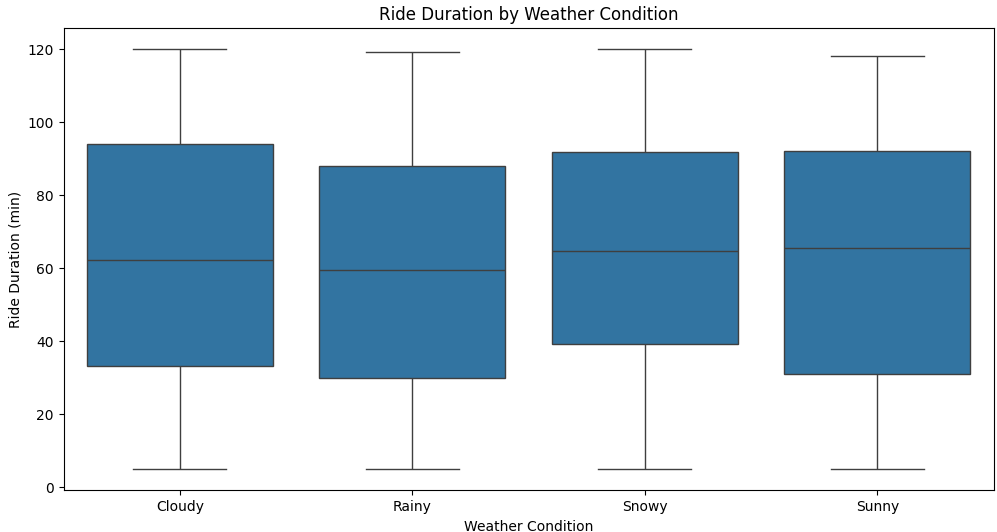
EDA was developed to summarize the key features as well as identifying trends in the data:

1. **Descriptive Statistics**:
   * **Rider Age**: The average age was approximately 40 years, later the most riders falling between 20 and 60 years.
   * **Ride Length**: Mean ride duration was 62 minutes, but durations varied, indicating diverse usage patterns in people.
   * **Bike Rental Cost ($)**: This had an average cost of $26, with a standard deviation of $9.39, showing variations based on different characteristics and possible environmental factors.
2. **Generic Visualizations**:
   * **Histograms** did not show any kind of skewness in this particular case.
   * **Boxplots** showed significant cost variations across all our bike models, with Electric Bikes costing more.
   * **Heatmaps** performed a strong correlation between **Rider Age** and **Bike Rental Cost**, as well as a moderate correlation with **Ride Duration**.
3. **Specific Charts:**
   * **Age Distribution of Riders**

*Figure 2: Age Distribution chart*

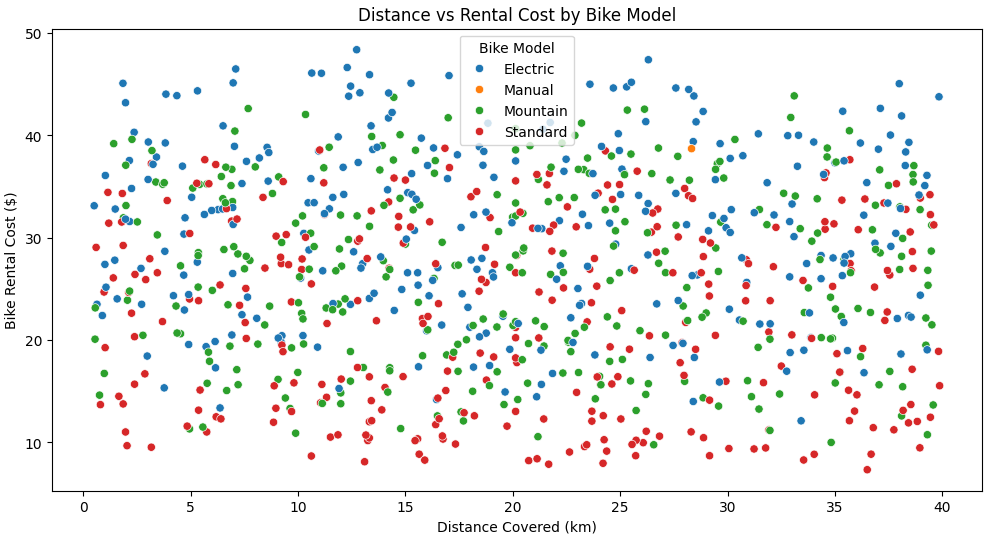
This confirms the Descriptive Analysis with having the more users in the age of 40.

* + **Ride Duration by Weather Condition**

*Figure 3: Ride Duration Boxplot*

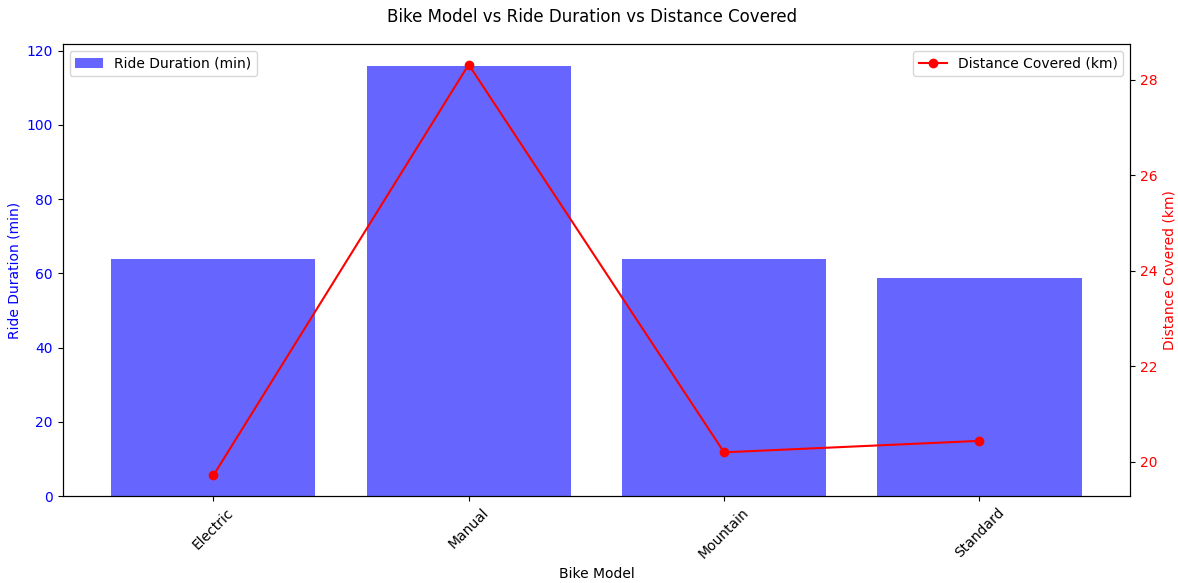
Surprisingly there is no much difference between weathers condition and Ride Duration, maybe this could tell us something in the future about the demographics of the users

* + **Distance Vs. Rental Cost by Bike Model**

*Figure 4: Distance Vs. Rental Cost Scatter Plot*

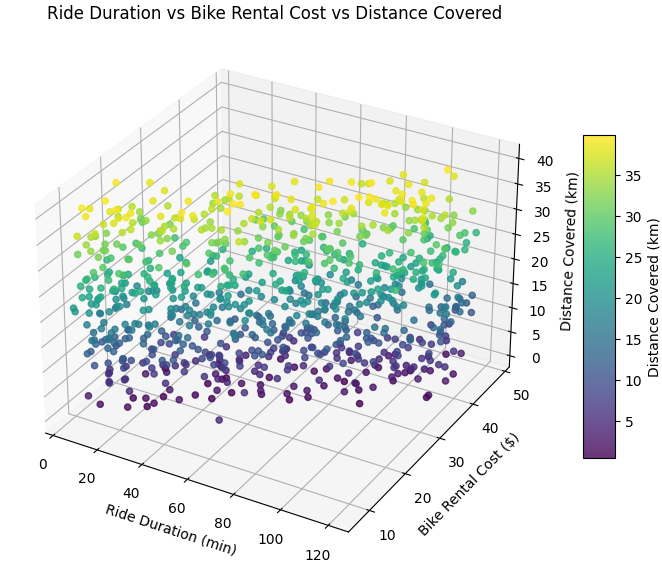
This one shows an interesting fact I learnt with this dataset. It almost does not matter the distance cost but the electric bike will generate more revenue than a standard bike. If we want to make the a profit we need to keep an eye on this chart

* + **Bike Model Vs. Ride Duration Vs. Distance Covered**

*Figure 5: Bike Model Vs. Ride Duration Vs. Distance Covered Chart*

So we find something interesting here, and it made sense with our previous chart, in our previous chart we analysed two extremes but we did not focus what was in the middle, in the middle are the manual bikes that seems to be the preferred ones overall.

* + **Ride Duration Vs. Bike Rental Cost Vs. Distance Covered**

*Figure 6: Bike Model Vs. Ride Duration Vs. Distance Covered Chart*

One of my favourite findings, we can see a strong correlation between Ride Duration with Bike Rental Cost and Distance Covered. So, what does this tells us? We need more people to ride more bikes for more time. Excursions plans, Premium monthly subscriptions, as well as not only bikes in the center of the city but in the borders so people stay more on the bike

**Feature Engineering**

To enhance the dataset, derived metrics were created:

**Numerical Encoding of Satisfaction**: **Rider Satisfaction** was encoded to High = 2, Medium = 1, and Low = 0, facilitating use in predictive modelling later on in the document.

**Dimensionality Reduction**

As we know high-dimensional data can lead to inefficiencies in analysis results. Two techniques were applied:

1. **Principal Component Analysis (PCA)**: We made the example with PCA for unsupervised learning
2. **Linear Discriminant Analysis (LDA)**: I made an example of LDA for regular supervised algorithms

With a clean and prepared dataset, I stepped into applying the basic statistical techniques for deeper exploration in the data.

**4. Statistical Techniques for Data Analytics**

Uncovering relationships within the data, validating hypotheses, and preparing insights in business decision-making, the statistical techniques are used for. To address the core business questions we created this section of descriptive and inferential analyses:

**Descriptive Analysis**

1. **Central Tendency Metrics**:
   * The mean rental cost was **$26**, with a median cost of **$26** and a mode of **$26**, basically the same, this reflects the affordability of these bike rentals.
   * Average rider age was **39 years**, followed by the majority of riders clustered between **20 and 60 years**, indicating a relatively young user base.
   * Average ride length was **62 minutes**,.
2. **Variation Metrics**:
   * **Rental Cost** showed a standard deviation of **$9.3**, indicating moderate pricing variability.

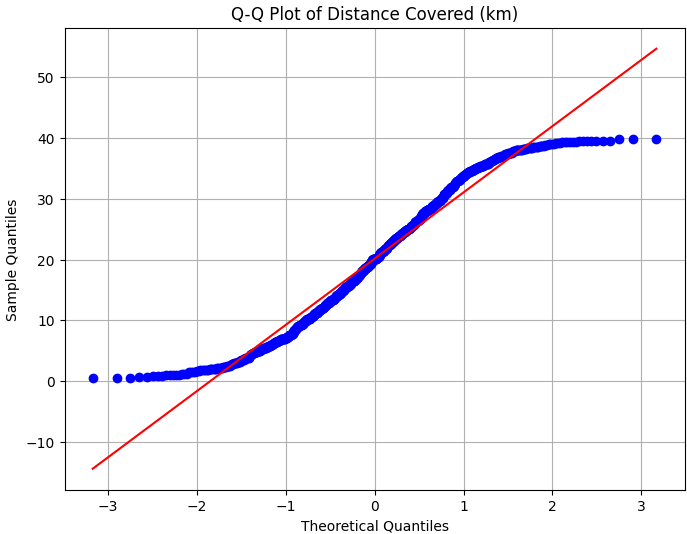
**Normality Tests**

The **Normality Test** was applied to check if there is normality in the numerical features. Results indicated:

* **All features** had non-normal distributions (p-values < 0.05).

**Statistical Visualizations**

1. **Scatterplots**:
   * Revealing a strong positive correlation in **Rental Cost** and **Distance Covered**. This showed me that longer rides tends to cost more in this dataset. This portrayed the importance of the Electric bikes in terms of making a profit, making them the most profitable of all. As well that showed us the Standard user that takes rides for the same amount of time paying less.
2. **Boxplots**:
   * Compared **Rider Age Vs. Rider Satisfaction,** this showed us the ones who were involved in giving feedback where the ones from 30 until 50 years of age, at the same time we learnt the level of satisfaction does not vary too much in age
3. **Heatmaps**:
   * We could see strong correlations among some numerical features. For example, **Distance Covered** and **Bike Rental Cost** were highly correlated, and **Ride Length** showing a moderate correlation.’
4. **Q-Q plot**: The feature with the better statistical results was the ‘Distance Covered (km)’ column, as we can see in this Q-Q plot.

*Figure 7: Q-Q Plot for numerical ‘Distance Covered (km)’ column*

**Confidence Intervals**

I stopped for a moment and allowed brainstormed ideas come to my way. I used confidence intervals to provided estimates of distance covered under different weather conditions:

* In **Sunny** conditions, the average ride length was **19km** (95% confidence).
* In **Rainy** conditions, the average ride length dropped to **17km** reflecting weather's impact on pricing.

**Hypothesis Testing**

I wanted to put to test this Rainy and Sunny theory, so what better than a T-test and our friend ANOVA to prove or disprove this point.

1. **T-tests** and **ANOVA**:
   * There was not significant different in both t-test and ANOVA between *sunny\_data* and *rainy\_data* failing to reject the null hypothesis.

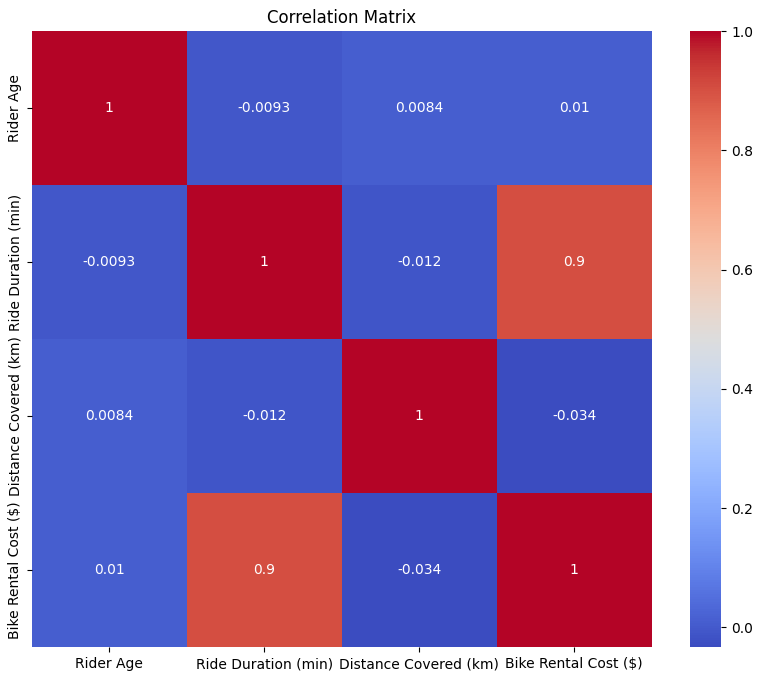
Maybe it was just pure chance.

**Correlation Analysis**

Correlation metrics helped prioritize predictive variables:

* **Rider Age** had a strong correlation with **Bike Rental Cost**, placing itself as the top predictor.
* **Linear Regression** confirmed this **Ride Duration** as the most influential variable, with an R² of **0.90**, making it explain the 90% of the variance in rental costs, confirming what previously the correlation map has showed.

*Figure 8: Correlation Analysis Matrix*



This statistical analysis provided a foundation for machine learning modeling, with this in our arsenal it is time to move on with the robustness of machine learning algorithms.

**5. Machine Learning**

Both supervised and unsupervised learning approaches were applied to answer key business so our machine learning phase focused on taking advantage of highly recognized advanced ML algorithms for printing interesting predictive insights and as well as segmenting customers appropriately.

**Supervised Learning**

1. **Model Selection**: Several models were employed for and regression tasks: Here, my goal was to predict **Bike Rental Costs ($)**.
   * **Linear Regression**: I used it trying to predict **Bike Rental Costs ($)**. The results were poor.
   * **Logistic Regression**: For predicting categorical outcomes.
   * **Random Forests**: For both classification and regression due to its interpretability and performance.
   * **Random Forest Regression**: Used to predict **Bike Rental Costs ($)** based on ride characteristics ‘Distance Covered (km)’. Results were not good.
2. **Hyperparameter Tuning**:
   * **GridSearchCV** was used to boost models (E.G.: the number of estimators in Random Forests).
3. **Cross-Validation**:
   * **10-fold cross-validation** were applied to reduce the risk of overfitting and helping the model generalizes as it should with the unseen data.

**Unsupervised Learning**

1. **K-Means Clustering**:
   * I used to let the unsupervised algorithm finds its perfect clusters and build them.
   * **Elbow Method** showed us three optimal clusters (k=3):
     + Cluster 0: Older riders preferring short rides on Electric Bikes, typically in urban or city areas.
     + Cluster 1: Mid aged people, riding a Standard bike, maybe for going to work.
     + Cluster 2: Younger, adventurous riders following excitement on Mountain Bikes choices.
2. **Results**:
   * Clustering insights were amazing, this will drastically change how future marketing campaigns will work on this enterprise.

**Evaluation**

Evaluation is critical to measure if the models are doing what they supposed to be doing, it is not only about predicting but predicting it right. This phase focuses on checking models performance, validating their insights or not, as well as making sure that the forwarded solutions aligns with business objectives.

**Supervised Learning Evaluation**

1. **Random Forest**:
   * This model has high errors (MSE = 98.4, RMSE = 9.92) in both of the splits.
   * Negative R2R values indicate poor fitting, model performs worse than predicting the mean.
2. **SVR**:
   * Marginally a better performance than our previous Random Forest with a slightly lower MSE (93.9–94.9) and RMSE (9.69–9.74).
   * Less negative R2R values but also fails.
3. **Ridge Regression**:
   * This one performed the best overall with the lowest MSE (93.1–93.6) and RMSE (9.65–9.67).
   * R2R values tells us it almost matches the mean but still struggles and fails.
4. **MAPE Analysis**:
   * High MAPE across all listed models (42–44%) shows signs of errors.

**Supervised Learning Conclusion**

All models presents poor predictive performance, what I think, due to abnormal data, as we study previously. Furthermore, Ridge Regression as the best of the group still requires further improvement through maybe feature engineering or tuning.

**Unsupervised Learning Evaluation**

1. **Silhouette Score: 0.372**
   * **0.372** shows a pass in clustering.
   * While it is true that the clusters are distinct, there is still the possibility of some that are not well clustered.
   * Score closer to **1** would show us a more distinct clustering of clusters, there is room for improvement.
2. **Calinski-Harabasz Score: 677.617**
   * High scores show that the clusters were well given.

**Unsupervised Learning Conclusion**

It is a pass for unsupervised learning because clustering is **successful** in this case.

**Real-world Aligmnet**

* **Operational Efficiency**:
  + This predictive models can guide a more dynamic fleet allocation based on the demand patterns predicted from weather and city.
* **Customer Satisfaction**:
  + Insights that we just gather helps in service offerings (E.G.: Special packages or promos for riding the bike in difficult weather)
* **Pricing Strategy**:
  + See the regression models, it will give understanding of user behaviour that help us to create a pricing strategy perfectly for each individual group.

**Limitations**

1. **Data Limitations**:
   * The dataset size (900 entries) was small for very modeling. Expanding the dataset could improve performance and clearer insights.
   * Missing data imputation may introduce biases, it does not matter if they are best practices it is still not the real data, so it can create noise in finding real patterns.
2. **Model Limitations**:
   * If time permitted we would have been able to find a solution of all Supervised algorithm performed poorly in this dataset.

**6. Deployment**

The deployment phase focuses on integrating what we have learnt in the analysis into the roadmap of the enterprise. If the company follows these instructions given in this report, they will secure growth in fleet management, customer satisfaction, and pricing.

**Marketing Campaigns**

1. **Promotions**:
   * Target campaigns:
     + Cluster 0: Discounts for city areas riders on elders using Electric Bikes.
     + Cluster 1: Monthly subscriptions for working aged people all around the globe.
     + Cluster 2: Adventure-themed campaigns promoting Mountain Bikes.
   * Weather-based promotions (e.g., discounts on rainy days) to incentivize rentals during lower demand periods.
2. **Loyalty Programs**:
   * Implement programs tailored to each segment, such as mileage rewards for frequent riders or age-based incentives for older customers.

The assurance that the analysis transitions from theory to practice is what we call deployment, that’s is why it is important to have in this report. My conclusions will summarize what we have been doing during in this project and how improvements are welcomed as well.

**7. Conclusion**

Guided by the CRISP-DM framework, this analysis , addressing the most crucial objectives, gives ready to use steps for the global bike rental company. Each step of the process has been structured to fit all their business goals, thanks to our data driven solutions.

By applying the recommendations this bike rental company can ensure a sustainable growth in a competitive market.

**Recommendations**

To improve the analysis:

1. **Enrich the Dataset**: Include variables such as time of ride, rental location, and promotions.

**Final thoughts**

The project places a foundation for understanding rider behavior relationship with profit. While it is true that the clustering analysis offered meaningful segmentation of riders, the regression models showed us the limitations of the current data. Future efforts should be making sure the quality and the amount of data is higher.

**8. References**

Artley, B. (2022). Unsupervised Learning: K-Means Clustering. [online] Medium. Available at: <https://towardsdatascience.com/unsupervised-learning-k-means-clustering-27416b95af27>.

Bhardwaj, A. (2020). Silhouette Coefficient : Validating clustering techniques. [online] Medium. Available at: <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>.

GeeksforGeeks (2024a). Difference between ttest and ANOVA. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/difference-between-t-test-and-anova/ [Accessed 26 Dec. 2024].

GeeksforGeeks (2024b). What is Ridge Regression? [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/what-is-ridge-regression/>.

Hotz, N. (2024). What Is CRISP DM? [online] Data Science Project Management. Available at: <https://www.datascience-pm.com/crisp-dm-2/>.

Jonatasv (2024). Metrics Evaluation: MSE, RMSE, MAE and MAPE. [online] Medium. Available at: <https://medium.com/@jonatasv/metrics-evaluation-mse-rmse-mae-and-mape-317cab85a26b>.

K, F. (2024). Page Restricted. [online] Medium.com. Available at: https://farshadabdulazeez.medium.com/understanding-silhouette-score-in-clustering-8aedc06ce9c4 [Accessed 26 Dec. 2024].

Malato, G. (2023). An Introduction to the Shapiro-Wilk Test for Normality | Built In. [online] builtin.com. Available at: <https://builtin.com/data-science/shapiro-wilk-test>.

MSc, F.E.O. (2023). Unsupervised Learning in R: Determination of Cluster Number. [online] Medium. Available at: <https://medium.com/@ozturkfemre/unsupervised-learning-determination-of-cluster-number-be8842cdb11>.

Petritis, B. (2018). t-test & ANOVA (Analysis of Variance). [online] Discovery in the Post-Genomic Age. Available at: <https://www.raybiotech.com/learning-center/t-test-anova/>.

Vungarala, S.K. (2023). PCA vs LDA — No more confusion! [online] Medium. Available at: <https://medium.com/@seshu8hachi/pca-vs-lda-no-more-confusion-fc21fb8d06e9>.

Walz, A. (2023). Top 10 Marketing Strategies for Bike Shops - Podium. [online] www.podium.com. Available at: <https://www.podium.com/article/top-10-marketing-strategies-for-bike-shops/>.

Yadav, A. (2024). Linear Regression vs Random Forest - Amit Yadav - Medium. [online] Medium. Available at: <https://medium.com/@amit25173/linear-regression-vs-random-forest-7288522be3aa>.

**9. GitHub Link**

<https://github.com/CCT-Dublin/ca2-Gabriel-studies>