Introduction:

Rental bikes are increasing in popularity and are being introduced into urban cities. Providing the city with a stable supply of rentals will become a concern if the demand keeps up. The research question for this project is to ensure supply is able to meet the future growing demand for bike rentals by predicting the hourly bike rentals. To solve this problem, tools and techniques such as R, Python, data cleaning, classification and regression trees, KNN are planned to be used.

Literature Review:
Several related papers and publications were reviewed to determine which factors may have a significant effect on when demand for rental bikes will be high.

A study done in Washington D.C. examined meteorological barriers that affect demand of rental bikes, and found that rainfall does not stop bike usage, seasons are a relatively unnecessary factor in demand, and very low humidity stops decreases demand. [1]

A publication using Random Forest Modeling examines the comparison between models trained with yearly and seasonal data to determine if hourly rental bike demand prediction can be more efficient by taking into account usage behaviour in different seasons and regions. [2]

A journal of transport geography investigated the effects of weather and calendar events on bike rentals and identified that both factors are negatively correlated with the daily demands for bike rentals, but not the same for time periods (high temperatures at 1am-7am are insignificant). [3]

A study using data mining techniques to predict demand for bike rentals in a metropolitan city determines that " weather data is shown to increase prediction accuracy" and looks to uncover correlation of weather data and the rental bike users count for a specific hour. [4]

Finally, a research paper uses multiple models like R2, RMSE, MAE and CV compared to RRF, CART, KNN and CIT, found that on all models, "Temperature or Hour was ranked as the most influential variable to predict the rental bike demand at each hour". [5]

Dataset:

The dataset used for this project can be found at https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand. The dataset consisted of 1 year's worth of bike rentals. The attributes used will include: Date, Number of bikes rented hourly, Hour, Temperature, Humidity, Rainfall, Snowfall, Seasons, Holiday, Functional Day(Closed/Open). Attributes that will not be used include: Windspeed,Visibility, Dew point temperature, Solar radiation, because the information is beyond the scope of this project.

Approach:

1. Data Processing/Cleaning
2. Exploratory Analysis
3. Feature Engineering
4. Visualisation
5. Modeling (Regression)

Step 1: Data Processing/Cleaning
The dataset is only for one year in a CSV format so there should not be too much to clean or process. May need to potentially change the type, format, checking for missing values, creating new columns (day of the week, month of the year).

Step 2: Exploratory Analysis
After the data is prepared and cleaned, exploratory analysis may begin. Searching for existing correlations in the data, and identification of attributes that will be used in the regression model later on.

Step 3: Feature Engineering
This step may not be required, but some attributes may need re-factoring to provide the regression model better inputs.

Step 4: Visualisation
Create plots to show all the correlations of the attributes. The plots will show a ranking of the attributes which are most correlated to the target, which will show what attributes will be needed to focus on.

Step 5:
This is a regression problem, because the target is to find the number of bikes rented per hour. Regression algorithms like decision tree, and linear multiple regression will be used to determine the algorithm with the best indicator.

References:
[1] Kumar, D. (2021). Meteorological barriers to bike rental demands: A case of Washington D.C. using NCA approach. Case Studies on Transport Policy, 9(2), 830-841. https://doi.org/10.1016/j.cstp.2021.04.002

[2] V E, S., & Cho, Y. (2020). Season wise bike sharing demand analysis using random forest algorithm. Computational Intelligence, https://doi.org/10.1111/coin.12287

[3] Kim, K. (2018). Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations. Journal of Transport Geography, 66, 309-320. https://doi.org/10.1016/j.jtrangeo.2018.01.001

[4] E, S. V., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353-366. https://doi.org/10.1016/j.comcom.2020.02.007

[5] V E, S., & Cho, Y. (2020). A rule-based model for seoul bike sharing demand prediction using weather data. European Journal of Remote Sensing, 53(sup1), 166-183. https://doi.org/10.1080/22797254.2020.1725789