Bike Sharing Assignment-based Subjective Questions

- Divya Srinivasan (I338215)

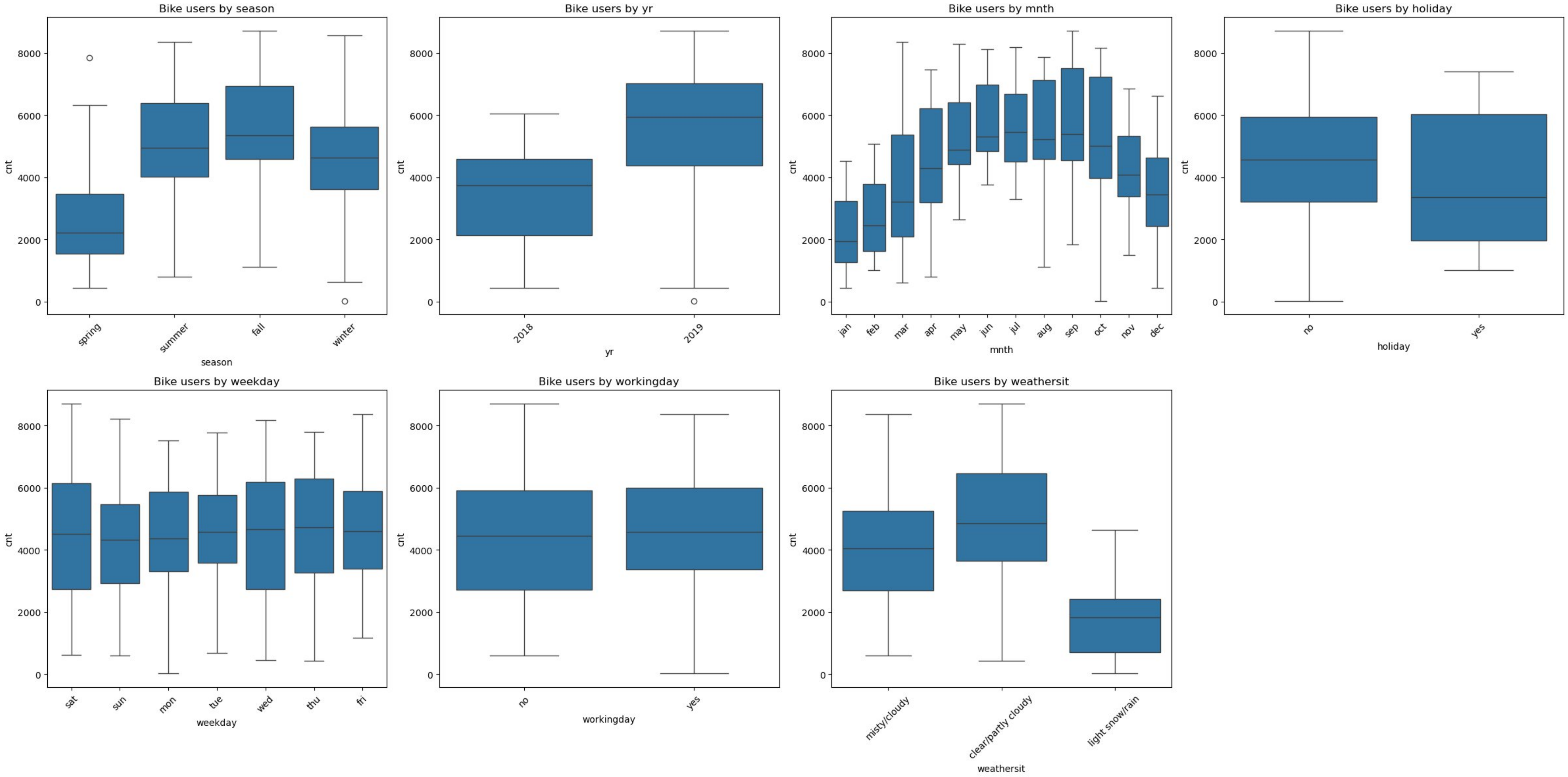
Q-1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

Please find the analysis of the effect of categorical variables on dependent variable (cnt) in the below table:

Categorical variable	Strength of influence of cnt	Analysis
season	Strong	'Fall’ season has high demand; ‘Spring’ has low demand;
weathersit	Very strong	Clear/partly cloudy is positively correlated
yr	Moderate	2019 had more demand than 2018
mnth	Moderate	April to Oct had more demand
weekday	Small to moderate	Monday to Friday had more demand than on Sunday
holiday	Small	Less demand on holidays
workingday	Small	More demand on working days

Plots that supports the above analysis comments



Q-2: Why is it important to use drop_first=True during dummy variable creation?

Answer:
It is important to use drop_first=True during dummy variable creation in Linear Regression to avoid multicollinearity due to dummy variable trap.

For example: Let’s take the categorical variable ‘season’ which has 4 values namely spring, summer, fall and winter. Now let’s create 4 dummy variables as shown below:

Without dropping first column (Incorrect approach)			
Spring	Summer	Fall	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

→

With dropping the first column (Correct approach)		
Summer	Fall	Winter
0	0	0
1	0	0
0	1	0
0	0	1

Notice that [Spring + Summer + Fall + Winter = 1], i.e. the four columns are perfectly correlated and one column could be predicted based on others. This prevents Linear Regression from computing unique coefficients.

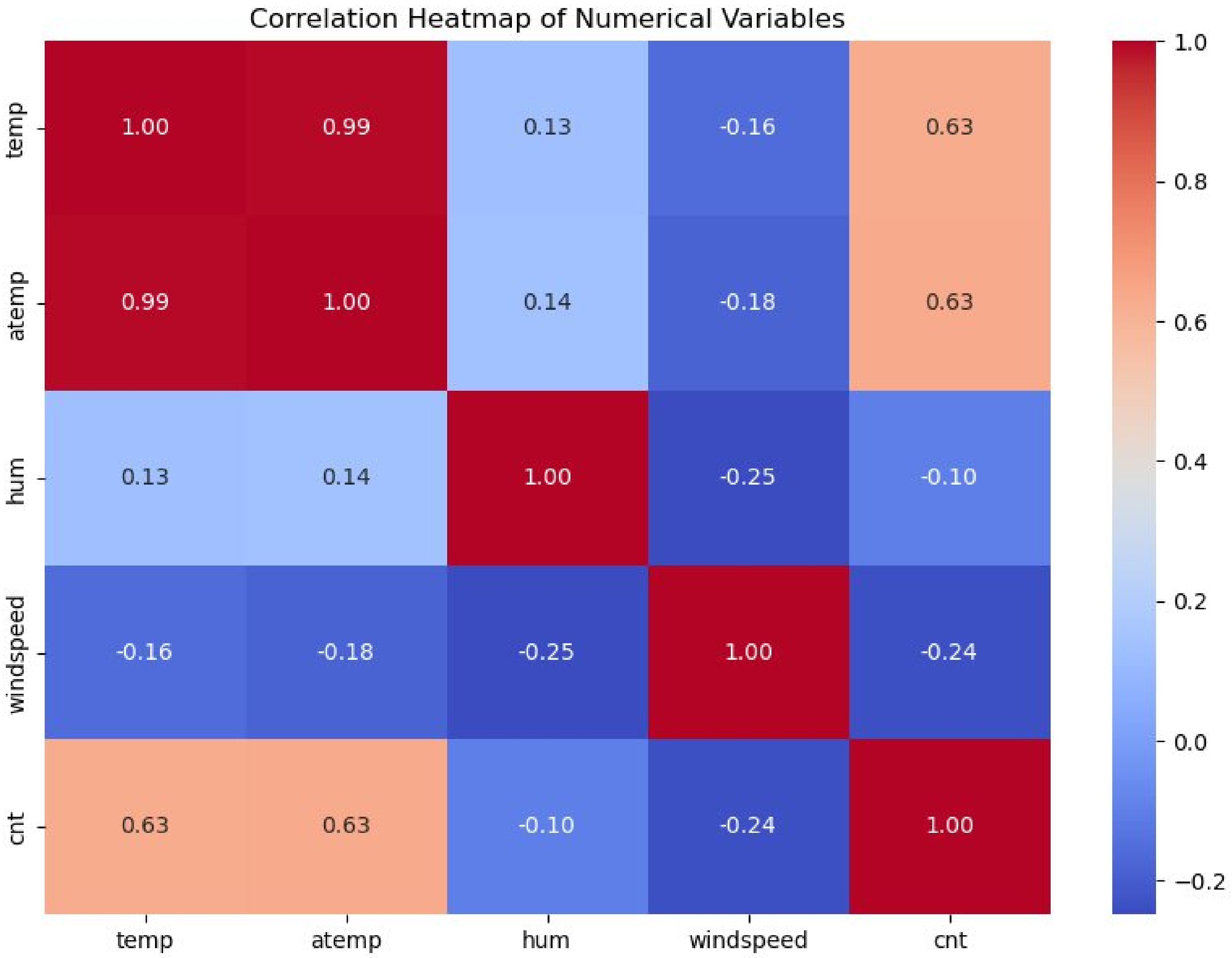
Q-3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:

The numeric variables ‘temp’ and ‘atemp’ are highly correlated with target variable ‘cnt’.

Since ‘temp’ and ‘atemp’ are almost similar, the variable ‘atemp’ is dropped after EDA. Thus, retaining ‘temp’.

Hence, ‘temp’ variable has the highest correlation with the target variable ‘cnt’.

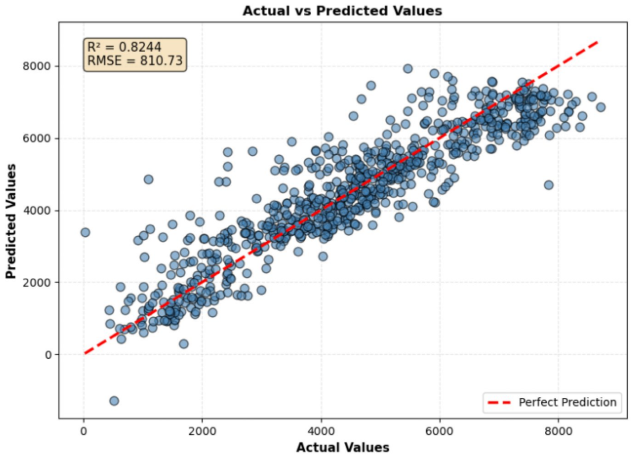


Correlation Heatmap of Numerical Variables

	temp	atemp	hum	windspeed	cnt
temp	1.00	0.99	0.13	-0.16	0.63
atemp	0.99	1.00	0.14	-0.18	0.63
hum	0.13	0.14	1.00	-0.25	-0.10
windspeed	-0.16	-0.18	-0.25	1.00	-0.24
cnt	0.63	0.63	-0.10	-0.24	1.00

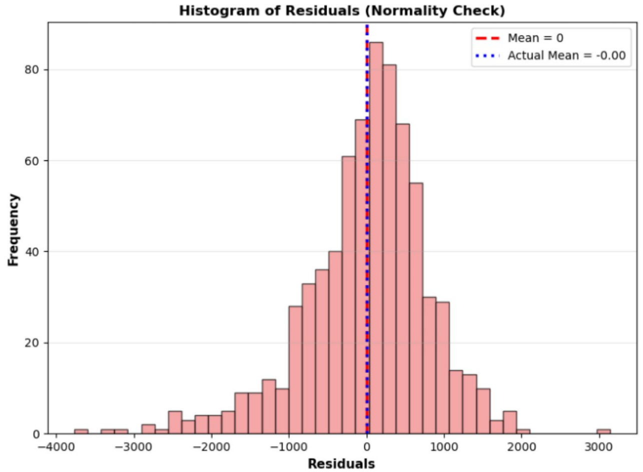
Q-4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:
The assumptions of Linear Regression were validated as follows:



Assumption-1: Linear relationship between X and Y

- Actual Vs Predicted scatter plot with Linear Regression line was plotted to verify this assumption.
- All the plotted points fit along the 45 degree Linear Regression line. Hence, they have linear relationship.



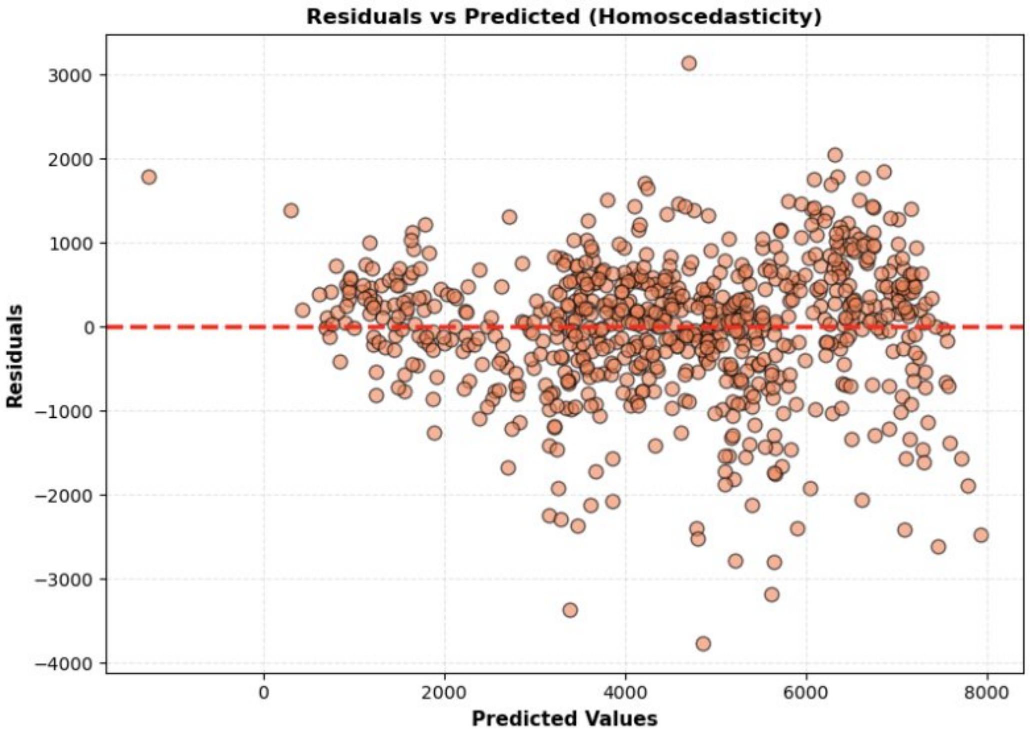
Assumption-2: Error terms are normally distributed (not X, Y)

- Histogram plot is bell-shaped and centered near zero indicating normally distributed errors.

Cont…

Assumption-3: Error terms have constant variance (homoscedasticity)

- From the Residual Vs Predicted scatter plot, it can be seen that there is no cone pattern.
- This indicates approximately constant variance of residuals.



Assumption-4: Error terms are independent of each other
- From the plots, it could be inferred that there could be slight correlation between error terms.

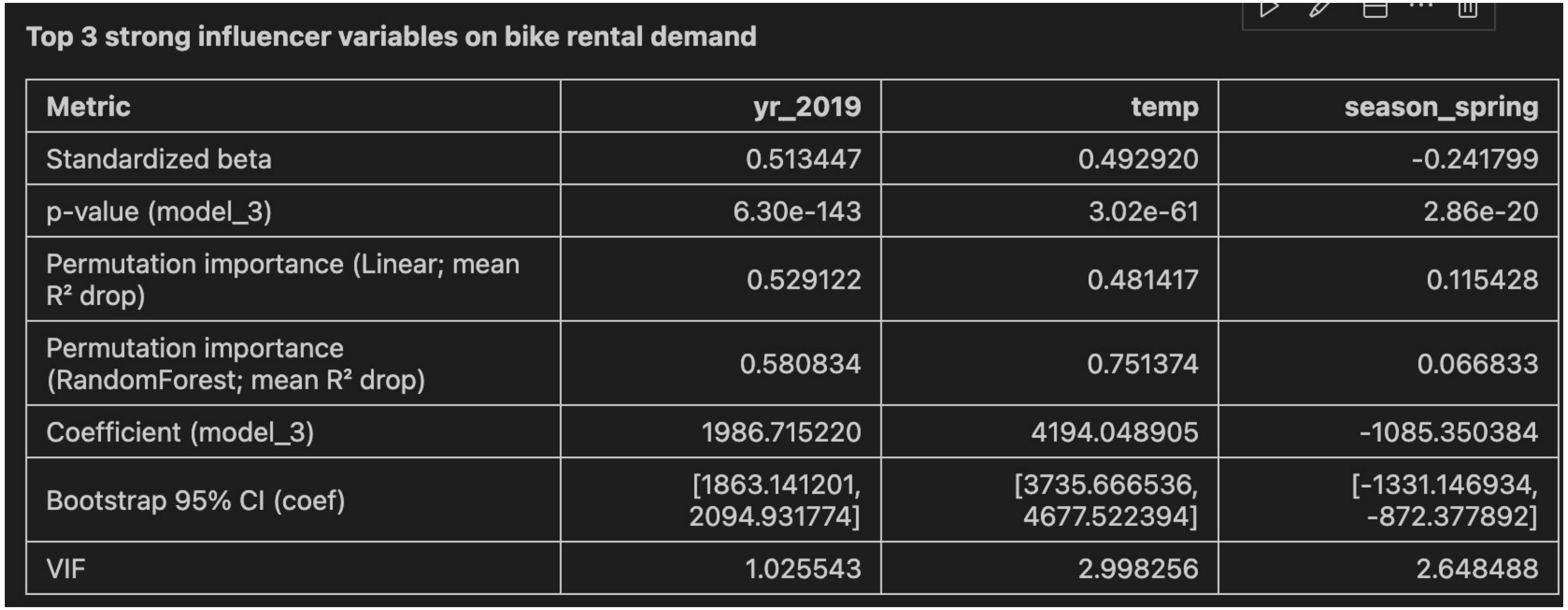
Q-5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

The top 4 features contributing significantly towards the shared bikes demand are:

1. Temp
2. Yr_2019
3. Season_spring

Detailed stats can be found in the below table (taken from Python Notebook):



Top 3 strong influencer variables on bike rental demand

Metric	yr_2019	temp	season_spring
Standardized beta	0.513447	0.492920	-0.241799
p-value (model_3)	6.30e-143	3.02e-61	2.86e-20
Permutation importance (Linear; mean R² drop)	0.529122	0.481417	0.115428
Permutation importance (RandomForest; mean R² drop)	0.580834	0.751374	0.066833
Coefficient (model_3)	1986.715220	4194.048905	-1085.350384
Bootstrap 95% CI (coef)	[1863.141201, 2094.931774]	[3735.666536, 4677.522394]	[-1331.146934, -872.377892]
VIF	1.025543	2.998256	2.648488

Bike Sharing
General Subjective Questions

Q-1: Explain the linear regression algorithm in detail.

Linear Regression:

- It is a Supervised Learning algorithm that predicts a dependent variable output based on one or more independent variables
- It fits a straight line that represents the linear relationship between the dependent and independent variables

Types of Linear Regression:

- Simple Linear Regression
- Multiple Linear Regression

Cont…

Simple Linear Regression:

Prediction of dependent variable based on one independent variable

Equation: y = mx + c, where

- y is the y-axes value
- x is the x-axis value
- m is the slope calculated as (co-variance of x,y) / (variance of x,y)
- c is the intercept of the Linear Regression line on y-axis

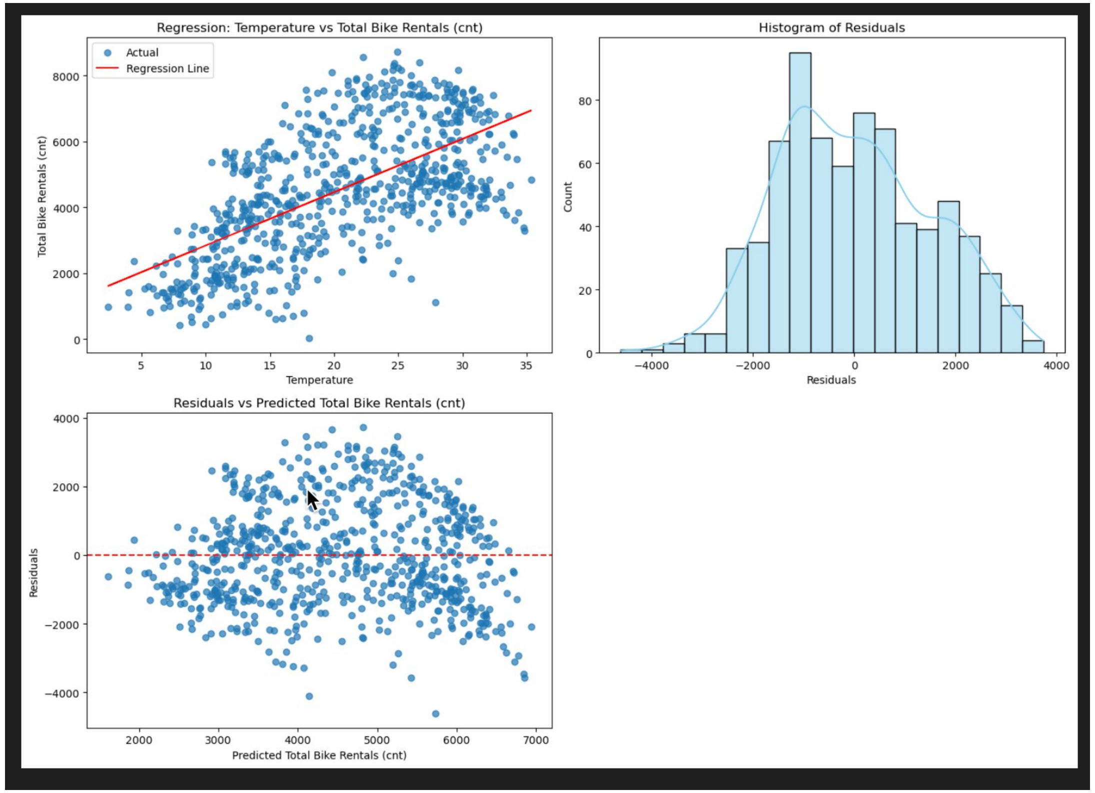
Strength of Linear Regression

Following techniques are used to access the best fit of Linear Regression:

- R-square or Coefficient of Determination
  - Higher the R-square, better the model fits
- Residual Standard Error (RSE)

Linear Regression assumptions are validated with the shown plots:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)



Cont…

Multiple Linear Regression

- Prediction of dependent variable based on multiple independent variable.

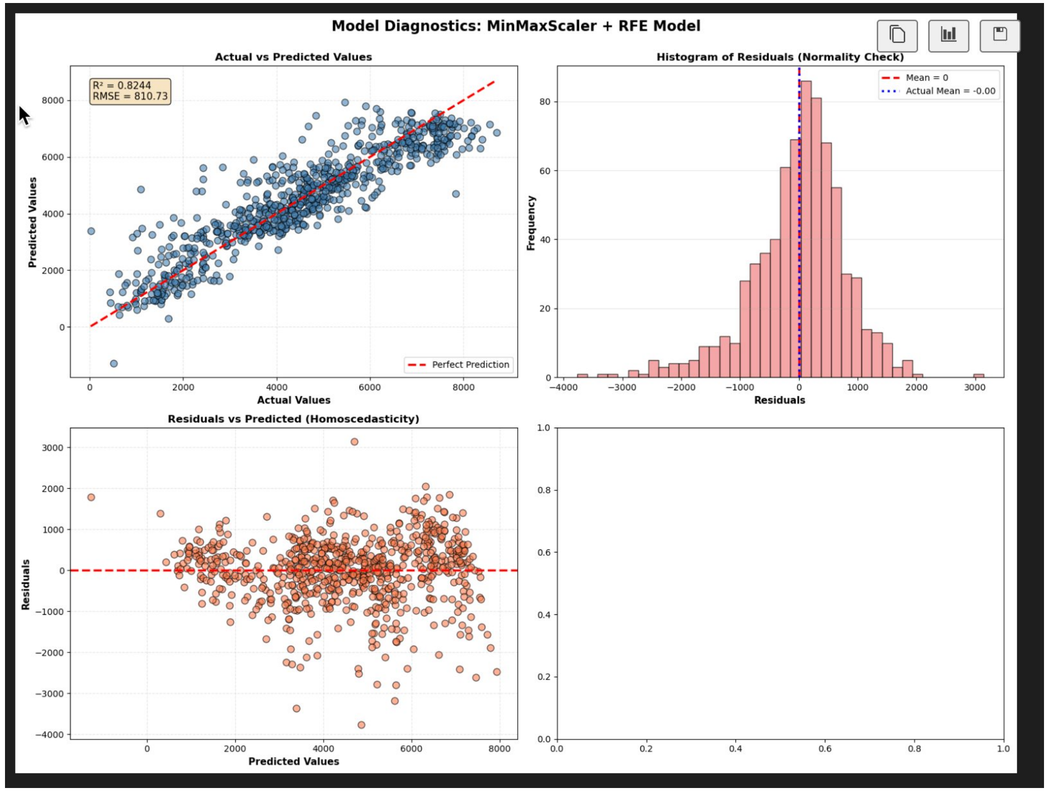
Feature Selection

- Model should not be too complex due to over fitting problem, so feature selection is required.
- Two types are there:
    - Standardizing using mean and standard deviation
    - MinMax scaling to range values between 0 and 1

Multicollinearity

- Highly correlated independent variables should be eliminated to avoid multi-collinearity.
- Categorical variables should be converted to numericals using dummy variables (should drop one column to avoid multicollinearity).

VIF (Variance Inflation Factor) – relationship of one independent variable with all others can be determined.
Adjusted R-square provides how best fit the model is.



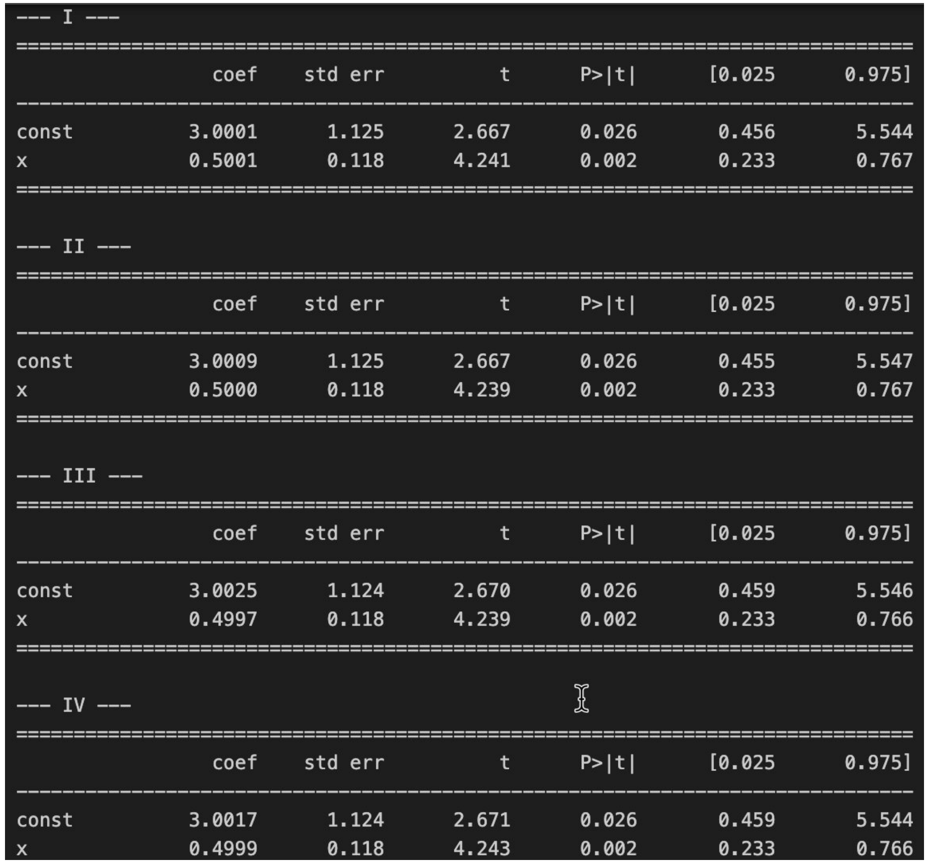
Q-2: Explain the Anscombe’s quartet in detail.

Anscombe’s quartet :

- It is a set of four datasets that have nearly identical summary statistics—same mean, variance, correlation, and regression line.
- However, when plotted, each dataset looks completely different.
- It shows that statistical measures alone can be misleading.
- The key lesson is that visualizing data is essential before drawing conclusions.

Example:

I took Anscombe’s dataset from seaborn library and calculated the key statistics. Results are as shown in the RHS screenshot. All stats are almost identical so its difficult to infer purely on this data.

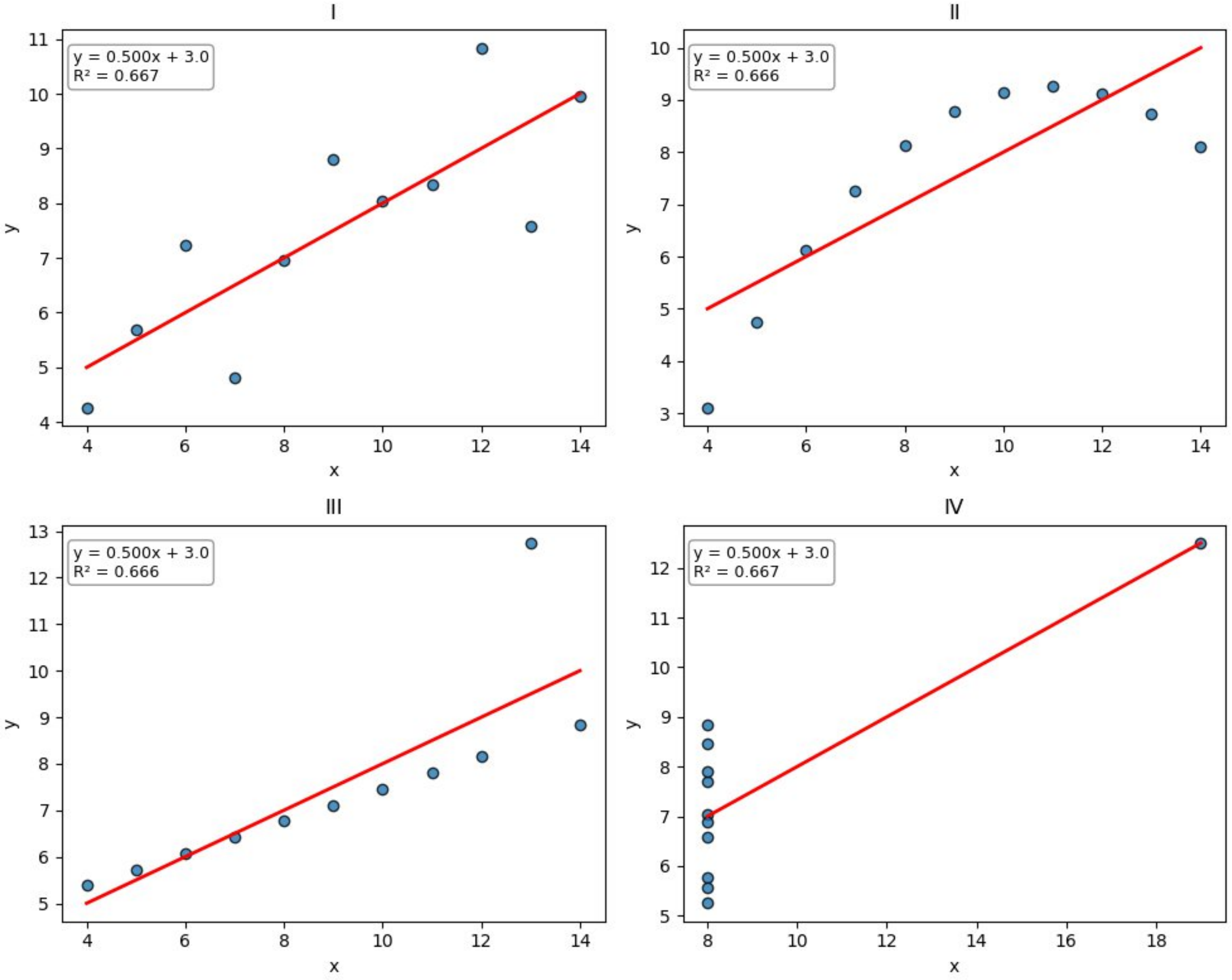


```
--- I ---
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0001      1.125      2.667      0.026       0.456       5.544
x              0.5001      0.118      4.241      0.002       0.233       0.767
==============================================================================

--- II ---
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0009      1.125      2.667      0.026       0.455       5.547
x              0.5000      0.118      4.239      0.002       0.233       0.767
==============================================================================

--- III ---
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0025      1.124      2.670      0.026       0.459       5.546
x              0.4997      0.118      4.239      0.002       0.233       0.766
==============================================================================

--- IV ---
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0017      1.124      2.671      0.026       0.459       5.544
x              0.4999      0.118      4.243      0.002       0.233       0.766
```

Cont…

Plots of Anscombe's dataset :
Plots of the 4 datasets reveal variations as can be seen:

- Dataset 1 has nearly linear relationship
- Dataset 2 has strong non-linear pattern
- Dataset 3 has linear pattern with one influential outlier
- Dataset 4 Almost no linear relationship, the one point is worst influencer



Q-3: What is Pearson’s R?

Pearson’s R:

- It is a statistical measure that tells you the strength and direction of a linear relationship between two variables.
- It ranges from –1 to +1
  - +1 - perfect positive linear relationship
  - –1 - perfect negative linear relationship
  - 0 - no linear relationship

- It measures how well the data fits a straight line.

Example:

- r = 0.85 → strong positive linear correlation
- r = –0.60 → moderate negative correlation
- r = 0.05 → almost no linear relationship

Q-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling:

- It is the process of transforming numerical features so they fit within a specific range or distribution. It ensures that all features contribute equally to a model.
- It helps models converge faster during training.
- It prevents bias toward features with larger units

Difference between Normalized Scaling vs Standardized Scaling:

Normalised scaling:

- Normalised scaling converts values into a fixed range, usually 0 to 1.
- Sensitive to outliers.
- Used when you need bounded values

Standardized Scaling:

- Converts data to have mean = 0 and standard deviation = 1.
- Not bounded.
- Less sensitive to outliers than Min-Max.

Q-5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A **VIF (Variance Inflation Factor)** becomes **infinite** when **multicollinearity** exists in the model.

Causes:

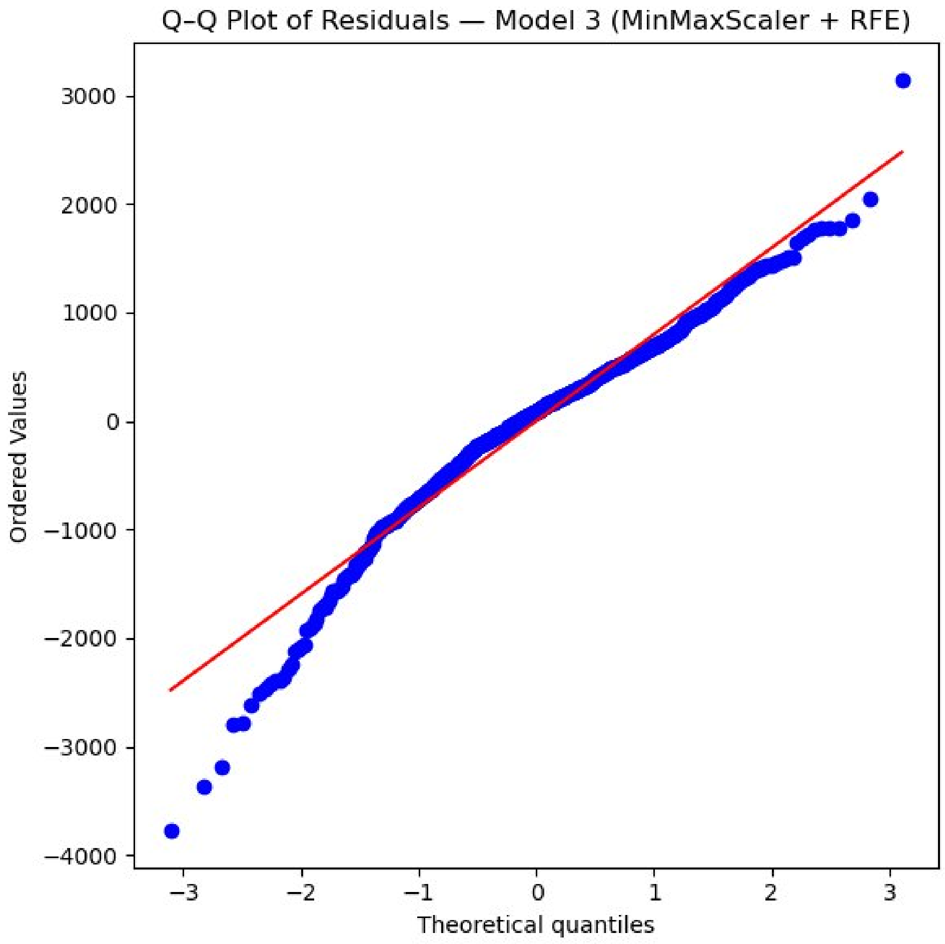
- Including **all dummy variables** (dummy variable trap).
- Having one feature that is exactly equal to another
  - Example: In bike sharing assignment, ’temp’ and ‘atemp’ are highly correlated. Hence, one has to be removed when building a model.
- Having a feature that is an exact sum of others
  - Example: ‘Registered’ and ‘Casual’ sum up to ‘cnt’ so the former two had to be removed

Q-6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plot:

- A Q–Q plot (Quantile–Quantile plot is used to check whether a dataset follows normal distribution.
- It compares the quantiles of your data against the quantiles of a normal distribution.
- If the data is normally distributed, the points in the plot will lie roughly along a straight 45° line.

Example from Bike sharing assignment for the best model selected can be seen in RHS



Usage in Linear Regression:

- Ensures the normality assumption of residuals.
- Confirms whether statistical inference (p-values, confidence intervals) is trustworthy.
- Helps diagnose model fit issues, outliers, or skewness.