Assignment-based Subjective Questions

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

A. There are many categorical variables in the data set like season, weather, holiday, working day, and year (yr). Inference of their effect on dependent variable i.e. Cnt is as:-

Season has strong significance on demand, fall (3) has highest demand as it's the favourable time to bike

Weather has negative influence on the demand, like rain, snow lowers the demand

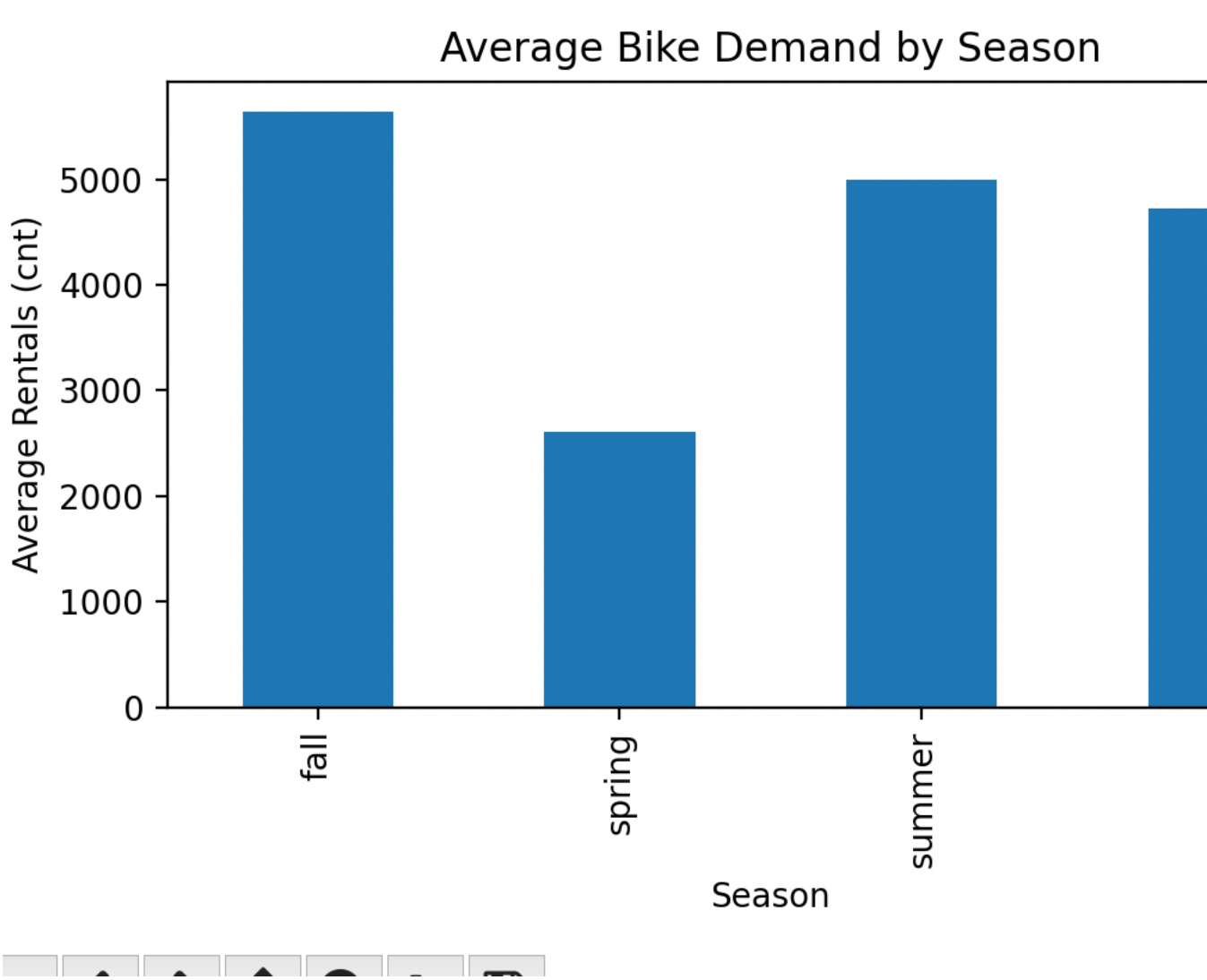
Year shows how the popularity or trend of biking is going as the year changes

So we can clearly say the seasonal categories show that demand is highest in fall and summer and low in spring, pointing that people prefer riding in moderate or pleasant weather.

The weather situation categories also reveal strong effects: clear or partly cloudy days lead to very high rentals, misty days lowers demand moderately, and days with rain or light snow show a sharp drop in usage. The year variable shows that demand increased significantly from 2018 to 2019, reflecting the growing popularity of bike-sharing services. Day-type categories show smaller but consistent effects—holidays generally show fewer rentals, while weekends and weekdays have only mild variations.

Together, these patterns suggest that categorical factors such as season, weather conditions, and year play a significant role in modelling customer behaviour and have an impactful influence on bike rental demand.

Plotting the graph also indicates the above



2. Why is it important to use drop_first=True during dummy variable creation?

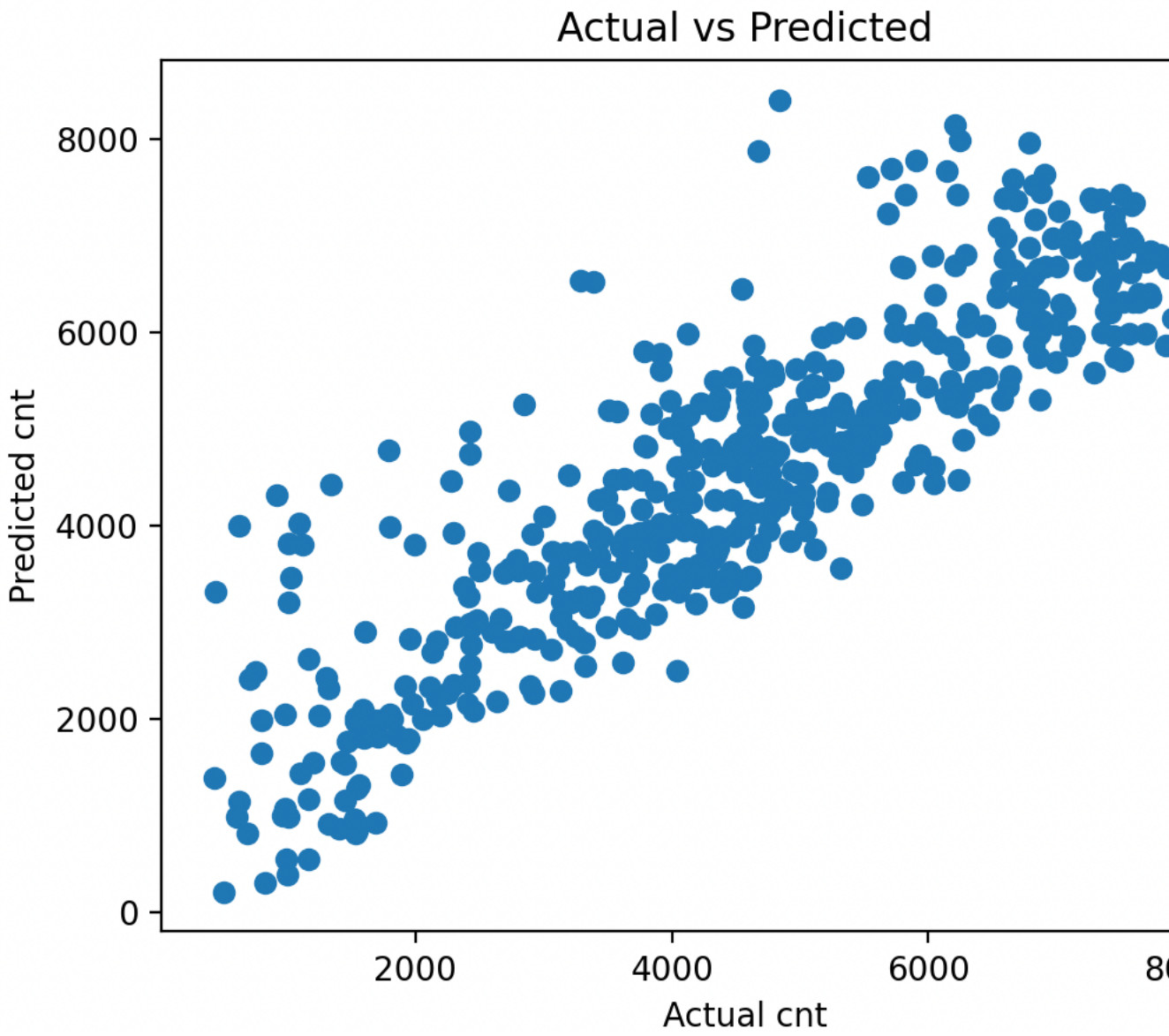
A. It avoids the dummy variable trap and helps making the linear regression model stable and mathematically valid

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

A. Temperature seems to be having the highest correlation with cnt.

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

A. Using seaborn library, I plotted a scatter plot between y_train_pred and y_train i.e actual cnt value and predicted cnt value. I observed it's a linear linear line and that tells me the assumption is not false.



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

A. Since the p-values for many variables are 0.00, I am looking at the coeff and as per the coefficient values- temperature, year and holiday are three variables that contribute significantly.

General Subjective Questions

1. Explain the linear regression algorithm in detail

A. Linear regression is a method to determine relationship between the target variable and featured variables also called independent variables.
If we plot the model on the graph, we will see the possible straight line through the data points
In equation terms:
$Y=b_0 + b_1x_1 + \ldots + b_nx_n$
where y -> predicted output
b_0 -> intercept
b_1 .. b_n -> are coefficients
x_1 .. x_n -> feature values

Linear regression indicates how well each feature contributes, if this contribution is positive or negative in relation to target variable and what will the expected output

To find the best fit line, linear regression uses OLS method which is Ordinary Least Squares.
Mathematical way to explain OLS will be –
$SSE= \sum(yi-yi\hat{})2$

Each coefficient tells how much the target variable changes for example if we see variable 'yr' then we can observe demand for bikes increased by 1942 in 2019 from 2018 year.

To ensure accuracy, linear regression assumes:

1. **Linearity**
2. **Normality of residuals**
3. **Homoscedasticity (constant variance)**
4. **Independence of errors**
5. **No multicollinearity**

2. Explain the Anscombe's quartet in detail.

A. Anscombe's quartet teaches that:

- Data must be **visualized**, not only summarized.
- Outliers, nonlinear patterns, and influential points can distort statistical metrics.
- Context and visualization are crucial to good statistical analysis.

Anscombe created four different datasets , each with

- Identical means of x& y
- Identical variance for x, y
- Identical correlation coefficient
- Identical Linear regression line(y=mx+c)

But when the graph was plotted, both look different and these were the observation:

Dataset 1 look liked it had a fairly linear relationship and data points were closer to regression line

Dataset 2 had a curved relationship, suggesting high correlation doesn't means a linear relationship

Dataset 3 indicated single **outlier in x** drastically affects the correlation and regression line

Dataset 4 showed nearly all points have the same x-value (vertical line), except one far outlying point, suggesting a single point can make it look like there is a linear relationship but in fact it isn't

3. What is Pearson's R?

3.A It Stands for Pearson correlation coefficient, is a statistic that measures the strength and direction of a linear relationship between two continuous variables.

Pearson's R quantifies how well the relationship between two variables *X* and *Y* can be described by a **straight line**.
It ranges from **–1 to +1**:

- **+1** → perfect positive linear correlation
- **–1** → perfect negative linear correlation
- **0** → no linear correlation

Key characteristics of Pearson's R are that it only captures linear relationships. A strong non-linear relationship can still give r ≈ 0.A single extreme value can drastically inflate or deflate r.
Pearson's R formally assumes:

- Both variables are continuous
- Both roughly follow a normal distribution
- The relationship is linear
- Variance is constant (homoscedasticity)

Usage -
When you want to measure linear relationship, Pearsons' R is helpful. When variable increases and other also tends to increase in straight line, Pearson's R is helpful.
When both the variables are continuous. Pearson's R assumes that we have quantitative numerical values that can vary smoothly. It relies on mean, variation and deviations from the mean.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of adjusting numerical data so that different features are on a similar scale. This is important because many machine-learning algorithms rely on distances or gradients, and features with large numerical ranges can dominate those with smaller ranges, slow down training, or create numerical instability.
There are two common types of scaling: normalization and standardization. Normalized scaling, also called min–max scaling, rescales values into a fixed range—usually between 0 and 1. It preserves the shape of the original distribution
but is highly sensitive to outliers because extreme values determine the minimum and maximum. Standardized scaling, or z-score

standardization, transforms data so that it has a mean of 0 and a standard deviation of 1.

In boom bikes dataset case, there are many numeric variables that have very different ranges like (yr- 0 or 1, weekday :0-6) These variables differ a lot on scale and while linear regression doesn't necessarily depend on scaling, but it can't be avoided as it is important to the use case. Because:

- Many features (numeric) dominate others
  Like weekday is between 0-6 while windspeed has ranges of 0-100, so if we don't scale the data, the variables with larger value will depict larger coefficients and influence the regression line
  So scaling in this scenarios ensures fairness and that each variable contributes fairly
- Scaling improves model stability
  Formula used by LR is $(X^T X)^{-1}$ So, if X contains variables with different ranges, the resultant matrix is not correct, causing incorrect p-values, unstable coefficients, warning issues, making the overall model unreliable
- Scaling also helps in understanding, interpreting the coefficients better
  If all the features are on same scale then we can confidently interpret that "Atemp is the strongest positive contributor" or "Humidity decreases bike demand"

  Without scaling, comparisons are misleading.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   VIF becomes infinite when one variable is perfectly predicted by one or more other variables.
   This means there is **perfect multicollinearity** in the dataset.
   This usually happens when:
   - Dummy variables are created but one category is not dropped (dummy variable trap).
   - Two columns are duplicates or one is a direct linear transformation of another.

- A feature becomes constant or has no variance due to preprocessing mistakes.

When this occurs, the R^2 value in the VIF formula becomes **1**, and as per the formula VIF = 1/0

So, an infinite VIF simply means a feature provides no unique information because it is completely explained by other features.

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

   A Q-Q plot, or Quantile–Quantile plot, is a visual tool used to check whether data follows a particular theoretical distribution, most commonly the normal distribution.

   In linear regression, it is used to assess whether the model's residuals are normally distributed. This matters because the validity of p-values, confidence intervals, and hypothesis tests in regression depends on the assumption of normally distributed residuals. If the residuals fall roughly along the diagonal line in the Q-Q plot, it suggests normality and indicates that the model assumptions are likely satisfied. If the points deviate significantly from the line, it signals issues such as skewness, heavy tails, or outliers, which may imply model misfit or the need for transformations or additional predictors.

   Q-Q plot helps in linear regression by:-

   Validating normality of residuals

   Detect skewness. If residual curve away from the line , upward curve would mean right skewness, downward curve would mean left skewness, S-shaped will mean outliers

   Indicates overall model fit and correctness

   If residuals deviate strongly from normality, it often means:

   - The model is mis-specified
   - Important predictors are missing
   - Non-linear relationships are not captured
   - Outliers are influencing results

-