Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Final model contains following highlighted categorical variable. We can see the a few categorical variables has positive coefficient i.e. these variables have positive effect on dependent variable and a few has negative coefficient i.e. these variables have negative effect on dependent variable.

Categorical variable	Coefficient
Yr	0.2480
Workingday	0.0575
Weekday_6	0.0513
Season_spring	-0.3198
Weathersit_2	-0.0870
Weathersit_3	-0.2925
Mnth_11	-0.1347
Mnth_12	-0.1131

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	200.9
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	7.67e-160
Time:	10:43:43	Log-Likelihood:	407.87
No. Observations:	510	AIC:	-795.7
Df Residuals:	500	BIC:	-753.4
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5181	0.018	29.394	0.000	0.483	0.553
yr	0.2480	0.010	25.377	0.000	0.229	0.267
workingday	0.0575	0.013	4.424	0.000	0.032	0.083
windspeed	-0.1494	0.027	-5.556	0.000	-0.202	-0.097
season_spring	-0.3198	0.012	-26.752	0.000	-0.343	-0.296
weathersit_2	-0.0870	0.010	-8.379	0.000	-0.107	-0.067
weathersit_3	-0.2925	0.030	-9.884	0.000	-0.351	-0.234
mnth_11	-0.1347	0.017	-7.848	0.000	-0.168	-0.101
mnth_12	-0.1131	0.019	-5.981	0.000	-0.150	-0.076
weekday_6	0.0513	0.017	3.012	0.003	0.018	0.085

Omnibus:	61.254	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.262
Skew:	-0.656	Prob(JB):	2.58e-30
Kurtosis:	5.166	Cond. No.	9.69

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

This ***avoids dummy variable trap*** (by dropping the first variable in the regression model).
i.e. the dummy variables created for categorical features are highly collinear and leading to multicollinearity issues in the regression model (if we don't drop first variable).

If we have n variables then n-1 variables are good enough to represent n variables. i.e. if we use all n variables then one of them become redundant as it can be derived from others. Hence, we always drop one variable.

Ex: In rental bikes demand assignment, following 4 seasons shall be represented by last 3 columns. In this case season_fall is a redundant column. This is true for all dummy variables. Hence, we need to drop the first column for all dummy variables.

	season_fall	season_spring	season_summer	season_winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

"atemp" has highest correlation with target variable "cnt".

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Perform residual analysis of train data, values shall be scattered around 0 without any specific pattern.
2. Features shall not be highly correlated with each other i.e. VIF shall be less than 5
3. The spread of residual shall be constant i.e. no funnel shaped pattern
4. The residuals shall be normally distributed
5. The P value shall be less than 0.05 for all features. Only these features shall have significance.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The following features contributing significantly towards the shared bikes demand

1. **yr:** This has positive coefficient as shared bike demand increased in 2019
2. **season_spring:** This has negative coefficient as bike demand decreases in spring
3. **weathersit_3:** This has negative coefficient as bike demand decreases in poor weather conditions (such as Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	200.9
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	7.67e-160
Time:	10:43:43	Log-Likelihood:	407.87
No. Observations:	510	AIC:	-795.7
Df Residuals:	500	BIC:	-753.4
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5181	0.018	29.394	0.000	0.483	0.553
yr	0.2480	0.010	25.377	0.000	0.229	0.267
workingday	0.0575	0.013	4.424	0.000	0.032	0.083
windspeed	-0.1494	0.027	-5.556	0.000	-0.202	-0.097
season_spring	-0.3198	0.012	-26.752	0.000	-0.343	-0.296
weathersit_2	-0.0870	0.010	-8.379	0.000	-0.107	-0.067
weathersit_3	-0.2925	0.030	-9.884	0.000	-0.351	-0.234
mnth_11	-0.1347	0.017	-7.848	0.000	-0.168	-0.101
mnth_12	-0.1131	0.019	-5.981	0.000	-0.150	-0.076
weekday_6	0.0513	0.017	3.012	0.003	0.018	0.085

Omnibus:	61.254	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.262
Skew:	-0.656	Prob(JB):	2.58e-30
Kurtosis:	5.166	Cond. No.	9.69

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression assumes linear relationship between dependent variable and the independent variables.

Types of linear regression:

1. **Simple linear regression:**
This shall contain only one independent variable
($y=\beta0+\beta1x$)

2. **Multiple linear regression:**
This shall contain multiple independent variables
($y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn$)

In this
y is independent variable
β0 is the intercept
β1, β2…βn are coefficients
x1, x2…xn are independent variables

Main objective is to find the best fitting line which minimizes the difference between observed values and predicted values. In this process we calculate β0, β1, β2, …, βn.

We can use ordinary least squares method to estimate these coefficients. OLS minimizes the sum of squared residuals (i.e. difference between observed and predicted values).

We evaluate this model based on R-squared, adjusted R-squared, mean square error and residual analysis.

Ex:
In the following we have calculated coefficients of independent variables and intercept. With help of these we can calculate the predicted value.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.783
Model:	OLS	Adj. R-squared:	0.780
Method:	Least Squares	F-statistic:	200.9
Date:	Sun, 24 Nov 2024	Prob (F-statistic):	7.67e-160
Time:	10:43:43	Log-Likelihood:	407.87
No. Observations:	510	AIC:	-795.7
Df Residuals:	500	BIC:	-753.4
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5181	0.018	29.394	0.000	0.483	0.553
yr	0.2480	0.010	25.377	0.000	0.229	0.267
workingday	0.0575	0.013	4.424	0.000	0.032	0.083
windspeed	-0.1494	0.027	-5.556	0.000	-0.202	-0.097
season_spring	-0.3198	0.012	-26.752	0.000	-0.343	-0.296
weathersit_2	-0.0870	0.010	-8.379	0.000	-0.107	-0.067
weathersit_3	-0.2925	0.030	-9.884	0.000	-0.351	-0.234
mnth_11	-0.1347	0.017	-7.848	0.000	-0.168	-0.101
mnth_12	-0.1131	0.019	-5.981	0.000	-0.150	-0.076
weekday_6	0.0513	0.017	3.012	0.003	0.018	0.085

Omnibus:	61.254	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.262
Skew:	-0.656	Prob(JB):	2.58e-30
Kurtosis:	5.166	Cond. No.	9.69

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet consists of 4 datasets with identical simple descriptive statistics, yet they appear different when they graphed. It emphasizes the importance of graphical representations in data analysis. Each data set has same mean, variance, correlation, and linear regression line. However, distinct visual patterns:

Dataset 1: Shows a linear relationship with no apparent outliers.

Dataset 2: Exhibits a perfect quadratic relationship with one significant outlier.

Dataset 3: Displays a linear relationship with one distinct outlier.

Dataset 4: Contains a cluster of points with one extreme outlier, forming a vertical pattern.

Anscombe's quartet shows how similar statistical properties can mask very different data

distributions. This highlights the needs for visual data analysis to anomalies and understand data better.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R also known as Pearson's correlation coefficient. This helps in measuring the linear relationship between two variables. Its value shall be between -1 to +1.

A value +1 indicates perfect positive relationship between variables
A value -1 indicates perfect negative relationship between variables
A value 0 indicate no linear relationship between variables

Formula is
$r=\sum(xi-\bar{x})(yi-\bar{y})/squareroot(\sum(xi-\bar{x})2\sum(yi-\bar{y})2)$

in this
xi, yi are individual data points
$\bar{x}$, $\bar{y}$ are the mean of X and Y

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling adjusts the range of feature values to ensure they fit within specific scale. This is important for machine learning models which has distance calculations (numeric values) etc.

Equal weights: Scaling ensures no single feature dominates due to its scale.
Improves model performance: enhances convergence speed in gradient descent
Consistency: Maintains data consistency across different features.

Types of scaling:

Normalized scaling (Min-Max):
This transforms the features to a range between 0 and 1.
Ex:
During the rental bike demand prediction model building, used the minmax scaler for numeric columns as shown below.

```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler()
```

```
# Apply scaler() to all the numeric columns except the dummy' variables, yr etc
num_vars=['atemp','hum','windspeed','cnt']
df_train[num_vars] = scaler.fit_transform(df_train[num_vars])
```

```
df_train.head()
```

	yr	holiday	workingday	atemp	hum	windspeed	cnt	season_spring
683	1	0	1	0.322150	0.639330	0.327101	0.605336	0
645	1	1	0	0.404998	0.731215	0.419004	0.609320	0
163	0	0	1	0.685963	0.509660	0.708724	0.554026	0
360	0	0	1	0.326273	0.785745	0.415925	0.088253	1
640	1	0	1	0.682653	0.817947	0.110593	0.862127	0

Standardized scaling:
Transforms features to a mean of 0 and standard deviation of 1.

This scaling is useful when the features have different units or when the data does not have a bounded range.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

Variance Inflation factor denotes multicollinearity among the predictor variables in a regression model. When VIF is infinite means one predictor is exact linear combination of the others.

Exact linear relationship: when one variable is perfectly predicted by others
Dummy variable trap: including all categories of categorical variable without dropping one
Repeated variable: The same variable included multiple times in different forms

To resolve this infinite VIF, requires removing or correcting the redundant predictor variable.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

With Q-Q plot (Quantile-Quantile plot), we can visually compare the quantiles of a dataset against a theoretical distribution (the normal distribution). This helps to determine, if the data follows the expected distribution.

In linear regression, Q-Q plots are used to check if residuals are normally distributed, by validating the assumption of normality. Points on the 45-degree reference line indicate normal distribution, while deviations suggest non-normality, mean it highlighting the potential issues. By revealing patterns such as skewness or kurtosis, Q-Q plots ensure the reliability and accuracy of the regression model. These are essential for figuring out model assumptions and improving model performance.

Ex: Q-Q plot of the rental bike demand prediction assignment

```
residuals = y_test - y_pred
sm.qqplot(residuals, line='45', fit=True)
plt.title('Q-Q plot of Residuals')
plt.show()
```