INTRODUCTION

you are a data professional working with historical marketing promotion data. You will use the data to run a one-way ANOVA and a post hoc ANOVA test. Then, you will communicate your results to stakeholders You have previously provided insights about how different promotion types affect sales; now stakeholders want to know if sales are significantly different among various TV and influencer promotion types.

To address this request, a one-way ANOVA test will enable you to determine if there is a statistically significant difference in sales among groups. This includes:

Using plots and descriptive statistics to select a categorical independent variable

Creating and fitting a linear regression model with the selected categorical independent variable

Checking model assumptions

Performing and interpreting a one-way ANOVA test

Comparing pairs of groups using an ANOVA post hoc test

Interpreting model outputs and communicating the results to nontechnical stakeholders

In [2]:
```
# Import libraries and packages.

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

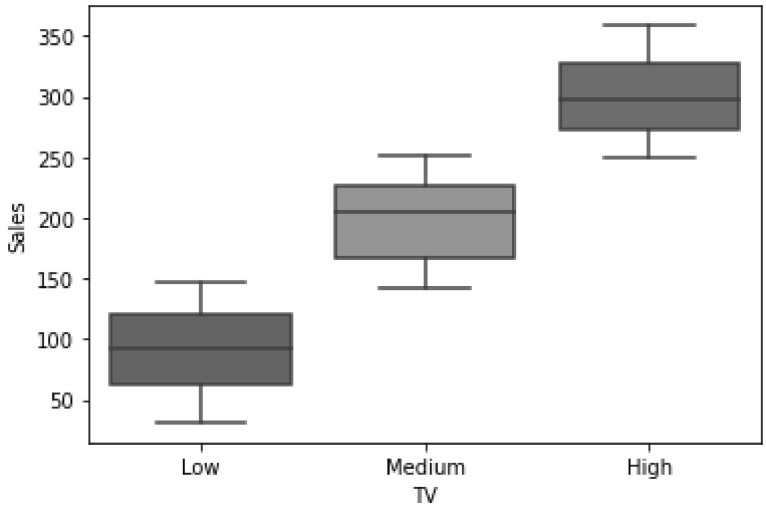
In [3]:
```
# Load the data.


data = pd.read_csv('Fk8EtlRPRyWPBLZUT-cl2Q_83304da42a7a433aa14111ee7a7c79f

# Display the first five rows.


data.head()
```

Out[3]:

	TV	Radio	Social Media	Influencer	Sales
0	Low	1.218354	1.270444	Micro	90.054222
1	Medium	14.949791	0.274451	Macro	222.741668
2	Low	10.377258	0.061984	Mega	102.774790
3	High	26.469274	7.070945	Micro	328.239378
4	High	36.876302	7.618605	Mega	351.807328

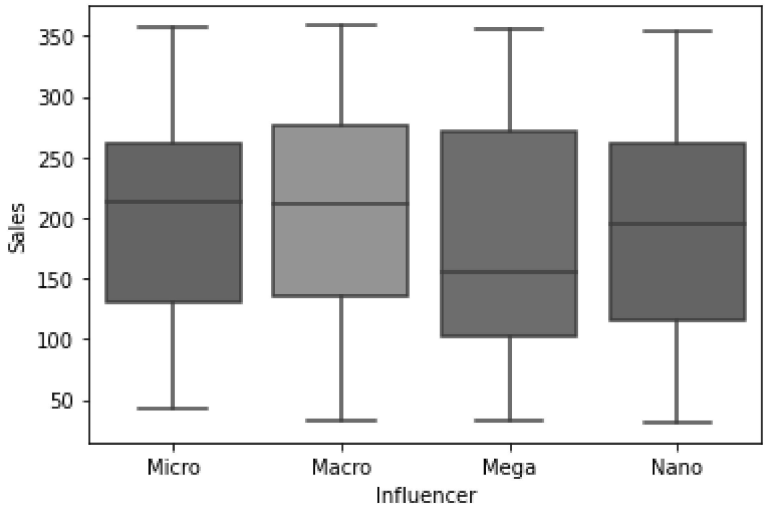
Data exploration

In [4]:
```
# Create a boxplot with TV and Sales.

sns.boxplot(x = "TV", y = "Sales", data = data);
```



from the above plot There is considerable variation in Sales across the TV groups. The significance of these differences can be tested with a one-way ANOVA

In [5]:
```
# Create a boxplot with Influencer and Sales.

sns.boxplot(x = "Influencer", y = "Sales", data = data);
```



There is some variation in Sales across the Influencer groups, but it may not be significant

In [7]:
```
data.isna().sum()
```

Out[7]:
```
TV             1
Radio          1
Social Media   0
Influencer     0
Sales          1
dtype: int64
```

In [9]:
```
data["Influencer"].value_counts()
```

Out[9]:
```
Mega     148
Nano     148
Micro    145
Macro    131
Name: Influencer, dtype: int64
```

In [13]:
```
1 data.groupby(["TV"])["Influencer"].value_counts()
```

Out[13]:
```
TV      Influencer
High    Macro         48
        Mega          43
        Micro         43
        Nano          43
Low     Mega          62
        Nano          53
        Micro         43
        Macro         39
Medium  Micro         59
        Nano          51
        Macro         44
        Mega          43
Name: Influencer, dtype: int64
```

In [14]:
```
1 # Drop rows that contain missing data and update the DataFrame.
2
3 data = data.dropna(axis=0)
4
5
6 # Confirm the data contain no missing values.
7
8 data.isnull().sum(axis=1)
```

Out[14]:
```
0      0
1      0
2      0
3      0
4      0
      ..
567    0
568    0
569    0
570    0
571    0
Length: 569, dtype: int64
```

Model building

In [15]:
```
# Define the OLS formula.

ols_formula = 'Sales ~ C(TV)'

# Create an OLS model.


OLS = ols(formula = ols_formula, data = data)

# Fit the model.

model = OLS.fit()

# Save the results summary.


model_results = model.summary()

# Display the model result

model_results
```

Out[15]:

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.874
Model:	OLS	**Adj. R-squared:**	0.874
Method:	Least Squares	**F-statistic:**	1971.
Date:	Fri, 02 Jun 2023	**Prob (F-statistic):**	8.81e-256
Time:	11:03:23	**Log-Likelihood:**	-2778.9
No. Observations:	569	**AIC:**	5564.
Df Residuals:	566	**BIC:**	5577.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	300.5296	2.417	124.360	0.000	295.783	305.276
C(TV)[T.Low]	-208.8133	3.329	-62.720	0.000	-215.353	-202.274
C(TV)[T.Medium]	-101.5061	3.325	-30.526	0.000	-108.038	-94.975

Omnibus:	450.714	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	35.763
Skew:	-0.044	**Prob(JB):**	1.71e-08
Kurtosis:	1.775	**Cond. No.**	3.86

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

note

TV was selected as the preceding analysis showed a strong relationship between the TV promotion budget and the average Sales. Influencer was not selected because it did not show a strong relationship to Sales in the analysis.

Check model assumptions: linearity

Because your model does not have any continuous independent variables, the linearity assumption is not required

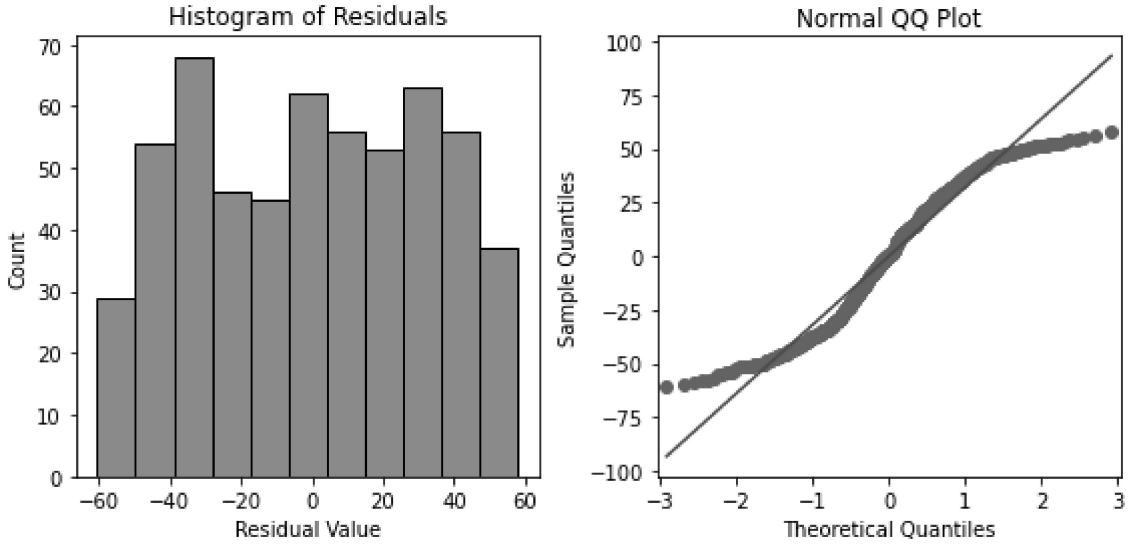
Check model assumptions: independent observation

The independent observation assumption states that each observation in the dataset is independent. As each marketing promotion (row) is independent from one another, the independence assumption is not violated.

Check model assumptions: normality assumption

Next, verify that the normality assumption is upheld for the model.

In [19]:
```
# Calculate the residuals.


residuals = model.resid

# Create a 1x2 plot figure.
fig, axes = plt.subplots(1, 2, figsize = (8,4))

# Create a histogram with the residuals.

sns.histplot(residuals, ax=axes[0])

# Set the x label of the residual plot.
axes[0].set_xlabel("Residual Value")

# Set the title of the residual plot.
axes[0].set_title("Histogram of Residuals")

# Create a QQ plot of the residuals.


sm.qqplot(residuals, line='s',ax = axes[1])

# Set the title of the QQ plot.
axes[1].set_title("Normal QQ Plot")

# Use matplotlib's tight_layout() function to add space between plots for
plt.tight_layout()

# Show the plot.
plt.show()
```

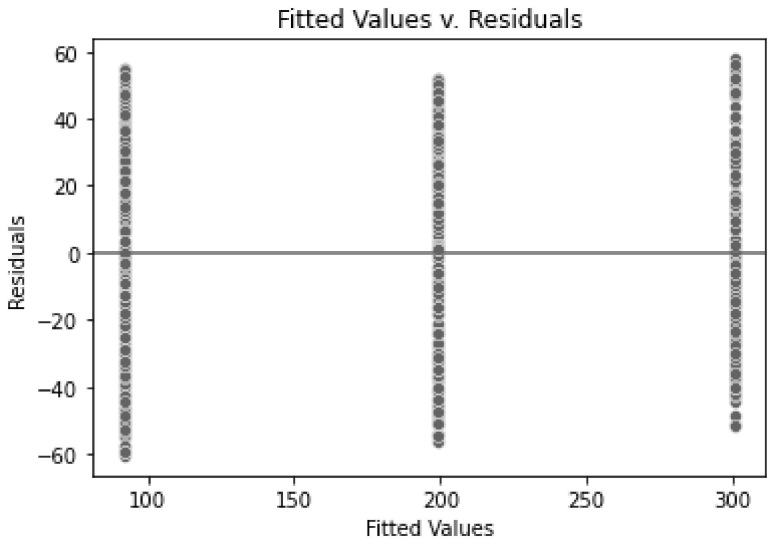


There is reasonable concern that the normality assumption is not met when TV is used as the independent variable predicting Sales. The normal q-q forms an 'S' that deviates off the red diagonal line, which is not desired behavior.

However, let us assume the normality assumption is met.

Check model assumptions: homoscedasticity

In [24]:
```
# Create a scatter plot with the fitted values from the model and the resi

fig = sns.scatterplot(x = model.fittedvalues, y = model.resid)

# Set the x axis label
fig.set_xlabel("Fitted Values")

# Set the y axis label
fig.set_ylabel("Residuals")

# Set the title
fig.set_title("Fitted Values v. Residuals")

# Add a line at y = 0 to visualize the variance of residuals above and bel


fig.axhline(0)

# Show the plot
plt.show()
```



The variance where there are fitted values is similarly distributed, validating that the constant variance assumption is met.

Results and evaluation

In [25]:
```
# Display the model results summary.

model_results
```

Out[25]:

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	0.874
Model:	OLS	**Adj. R-squared:**	0.874
Method:	Least Squares	**F-statistic:**	1971.
Date:	Fri, 02 Jun 2023	**Prob (F-statistic):**	8.81e-256
Time:	11:03:23	**Log-Likelihood:**	-2778.9
No. Observations:	569	**AIC:**	5564.
Df Residuals:	566	**BIC:**	5577.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	300.5296	2.417	124.360	0.000	295.783	305.276
C(TV)[T.Low]	-208.8133	3.329	-62.720	0.000	-215.353	-202.274
C(TV)[T.Medium]	-101.5061	3.325	-30.526	0.000	-108.038	-94.975

Omnibus:	450.714	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	35.763
Skew:	-0.044	**Prob(JB):**	1.71e-08
Kurtosis:	1.775	**Cond. No.**	3.86

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Using TV as the independent variable results in a linear regression model with R2=0.871. In other words, the model explains 87.1% of the variation in Sales. This makes the model an effective predictor of Sales.

The default TV category for the model is High, because there are coefficients for the other two TV categories, Medium and Low. According to the model, Sales with a Medium or Low TV category are lower on average than Sales with a High TV category. For example, the model predicts that a Low TV promotion would be 209.8691 (in millions of dollars) lower in Sales on average than a High TV promotion.

The p-value for all coefficients is 0.000, meaning all coefficients are statistically significant at p=0.05. The 95% confidence intervals for each coefficient should be reported when presenting results to stakeholders. For instance, there is a 95% chance the interval [−216.535,−203.203]

contains the true parameter of the slope of &TV Low, which is the estimated difference in

Given how accurate TV was as a predictor, the model could be improved with a more granular view of the TV promotions, such as additional categories or the actual TV promotion budgets. Further, additional variables, such as the location of the marketing campaign or the time of year, may increase model accuracy

Perform a one-way ANOVA test

In [26]:
```
1 sm.stats.anova_lm(model, typ = 1)
```

Out[26]:

	df	sum_sq	mean_sq	F	PR(>F)
C(TV)	2.0	4.052692e+06	2.026346e+06	1971.455737	8.805550e-256
Residual	566.0	5.817589e+05	1.027843e+03	NaN	NaN

The null hypothesis is that there is no difference in Sales based on the TV promotion budget.

The alternative hypothesis is that there is a difference in Sales based on the TV promotion budget.

The F-test statistic is 1917.75 and the p-value is $1.38*10-253$ (i.e., very small). Because the p-value is less than 0.05, you would reject the null hypothesis that there is no difference in Sales based on the TV promotion budget.

The results of the one-way ANOVA test indicate that you can reject the null hypothesis in favor of the alternative hypothesis. There is a statistically significant difference in Sales among TV groups.

Perform an ANOVA post hoc test

In [27]:
```
1 # Perform the Tukey's HSD post hoc test.
2
3 tukey_oneway = pairwise_tukeyhsd(endog = data["Sales"], groups = data["TV"
4
5 # Display the results
6 tukey_oneway.summary()
```

Out[27]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
High	Low	-208.8133	-0.0	-216.6367	-200.9898	True
High	Medium	-101.5061	-0.0	-109.3202	-93.6921	True
Low	Medium	107.3072	-0.0	99.7066	114.9077	True

The first row, which compares the High and Low TV groups, indicates that you can reject the null hypothesis that there is no significant difference between the Sales of these two groups.

You can also reject the null hypotheses for the two other pairwise comparisons that compare High to Medium and Low to Medium.

A post hoc test was conducted to determine which TV groups are different and how many are different from each other. This provides more detail than the one-way ANOVA results, which can at most determine that at least one group is different. Further, using the Tukey HSD controls for the increasing probability of incorrectly rejecting a null hypothesis from peforming multiple tests.

The results were that Sales is not the same between any pair of TV groups.

Note

Box-plots are a helpful tool for visualizing the distribution of a variable across groups. One-way ANOVA can be used to determine if there are significant differences among the means of three or more groups. ANOVA post hoc tests provide a more detailed view of the pairwise differences between groups.

Findings

High TV promotion budgets result in significantly more sales than both medium and low TV promotion budgets. Medium TV promotion budgets result in significantly more sales than low TV promotion budgets.

Specifically, following are estimates for the average difference between each pair of TV promotions:

1. Estimated average difference between High and Low TV promotions: $209.87 million (with 95% confidence that the exact value for this average difference is between 201.89 and 217.84 million dollars).
2. Estimated average difference between High and Medium TV promotions: $105.50 million (with 95% confidence that the exact value for this average difference is between 97.56 and 113.43 million dollars).
3. Estimated average difference between Medium and Low TV promotions: $104.37 million (with 95% confidence that the exact value for this average difference is between 96.83 and 111.92 million dollars).

The linear regression model estimating Sales from TV had an R-squared of $0.871, making it a fairly accurate estimator. The model showed a statistically significant relationship between the TV promotion budget and Sales. The model estimated the following relationships:

Using a high TV promotion budget instead of a medium TV promotion budget increased sales by 105.4952 million dollars (95% CI - 98.859, 112.131 million dollars).

Using a high TV promotion budget instead of a low TV promotion budget increased sales by 209.8691 million dollars (95% CI - 203.203 million, 216.535 million dollars).

Conclusion

1. influencer promotion type is not a good sales predictor
2. TV promotion type is a good sales prdictor specifically the high type

In [ ]:
```
1
```