My_Project

November 28, 2023

This project considers data in a Microsoft Excel File. The data consists of sales for the four quarters of 2022 by 153 salespersons employed by a multi-sectoral organization. The variables include age, sex, marital status, education level, department, number of assistants, experience, marketing budget(KES 1,000,000), appraisal score and their roles for the four quarters of 2022. The sales are in KES 1,000,000. I used python to perform the following data analysis tasks: 1. Use appropriate graphs to present the data 2. Perform the relevant descriptive analysis 3. Test the hypothesis of equality of proportions of female and male salespersons in each department and the entire organization 4. To test whether there's a significant relationship between marital status and education level 5. To test whether there's a significant relationship between education level and deprtment 6. Compare the mean quarterly and annual sales by age, sex, marital status, education level and department. For age, create a category for young employees as those aged 35 years and below, old otherwise 7. Figure out whether there's a quarter where the sales are significantly different from others 8. Repeat 6) but using the non-parametric approach 9. Repeat 7) but using the non-parametric approach 10. Fit multiple linear regression models for quarterly and annual sales on age, number of assistants, experience, marketing budget and appraisal score. Evaluate whether the models are a good fit and test the significance of each of the independent variables

[1]:
```
#Set up the coding environment by importing the necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from scipy.stats import ttest_ind, chi2_contingency, f_oneway, kruskal
from statsmodels.stats.multicomp import pairwise_tukeyhsd
import statsmodels.api as sm
```

[2]:
```
#ignore warnings
import warnings
warnings.filterwarnings('ignore')
```

[3]:
```
#Load the data from the excel file
xlsx_file_path = 'C:\\Users\\ADMIN\Downloads\Sales_data.xlsx'
Sales_data = pd.read_excel(xlsx_file_path, sheet_name='Sheet1')
Sales_data.head()
```

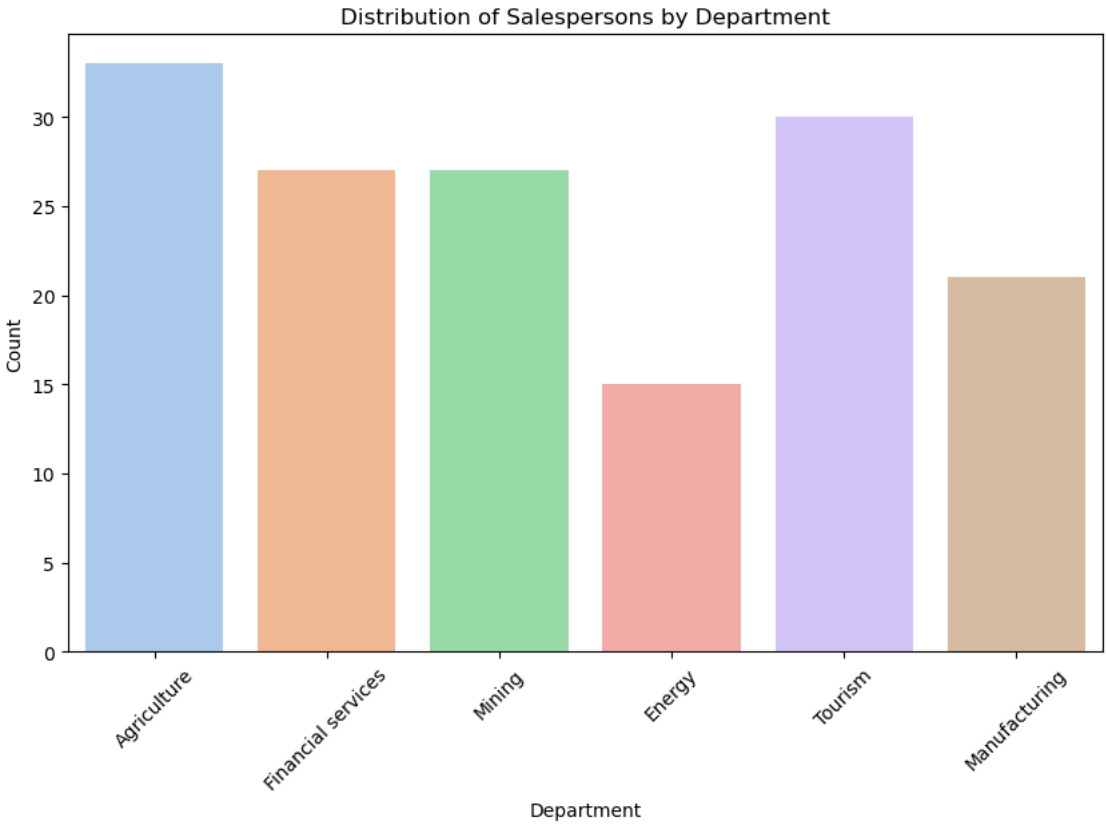
[3]:
```
  Personel Number     Sex  Age Marital Status Education Level  \
0           P0010  Female   57        Married     Certificate
```

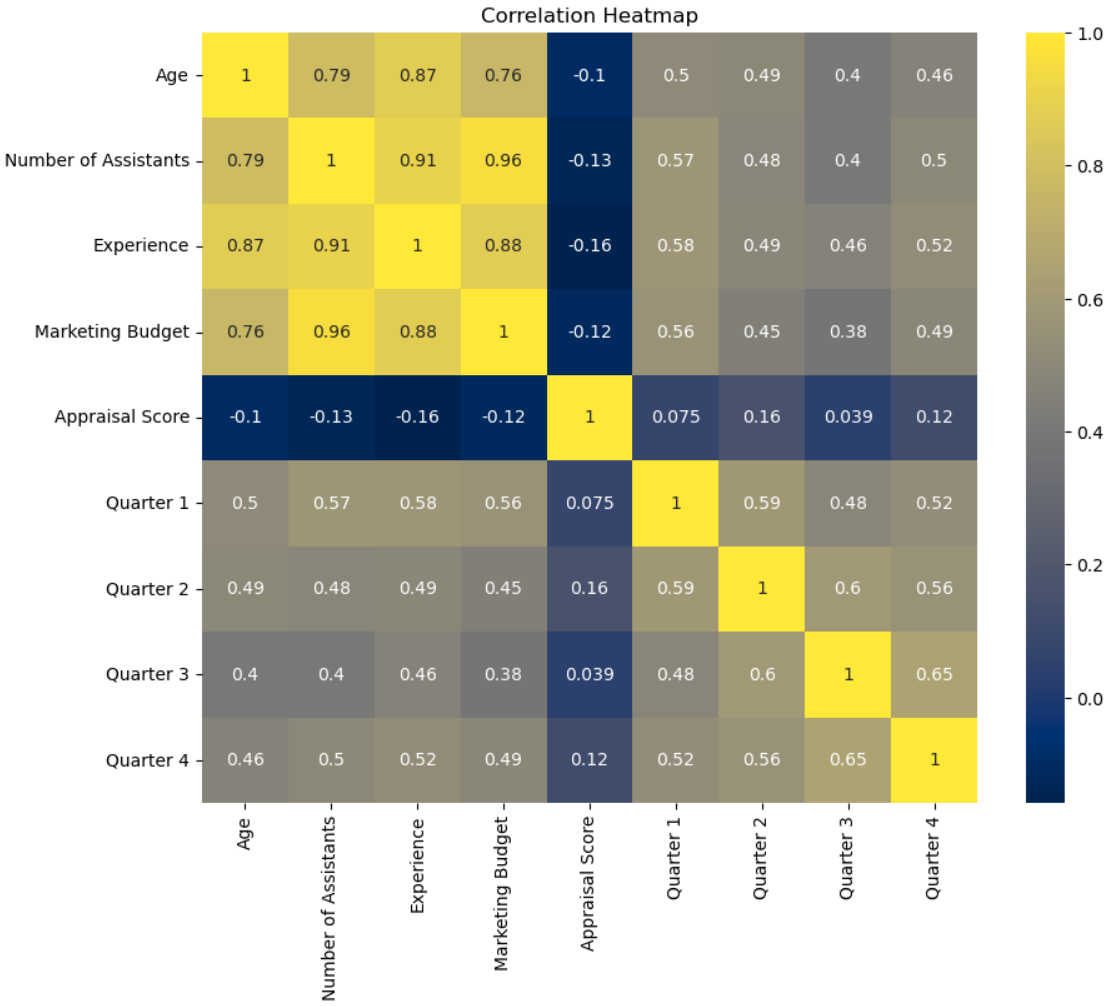
```
1           P0012    Male   37        Married     Certificate
2           P0018    Male   54       Divorced         Diploma
3           P0020    Male   59        Married     Certificate
4           P0023  Female   56        Married   Post graduate

           Department  Number of Assistants  Experience  Marketing Budget  \
0         Agriculture                    14          14              2.02
1  Financial services                    10           9              1.09
2              Mining                    25          20              3.34
3  Financial services                    26          19              3.45
4  Financial services                    31          24              4.60

   Appraisal Score  Quarter 1  Quarter 2  Quarter 3  Quarter 4
0             88.6      18.19      11.04      16.72      18.74
1             66.9      12.12      16.63      16.91      15.81
2             72.4      23.06      13.59      11.20      15.18
3             71.1      15.00      19.95      16.54      14.98
4             76.9      14.17      12.37      21.02      23.96
```

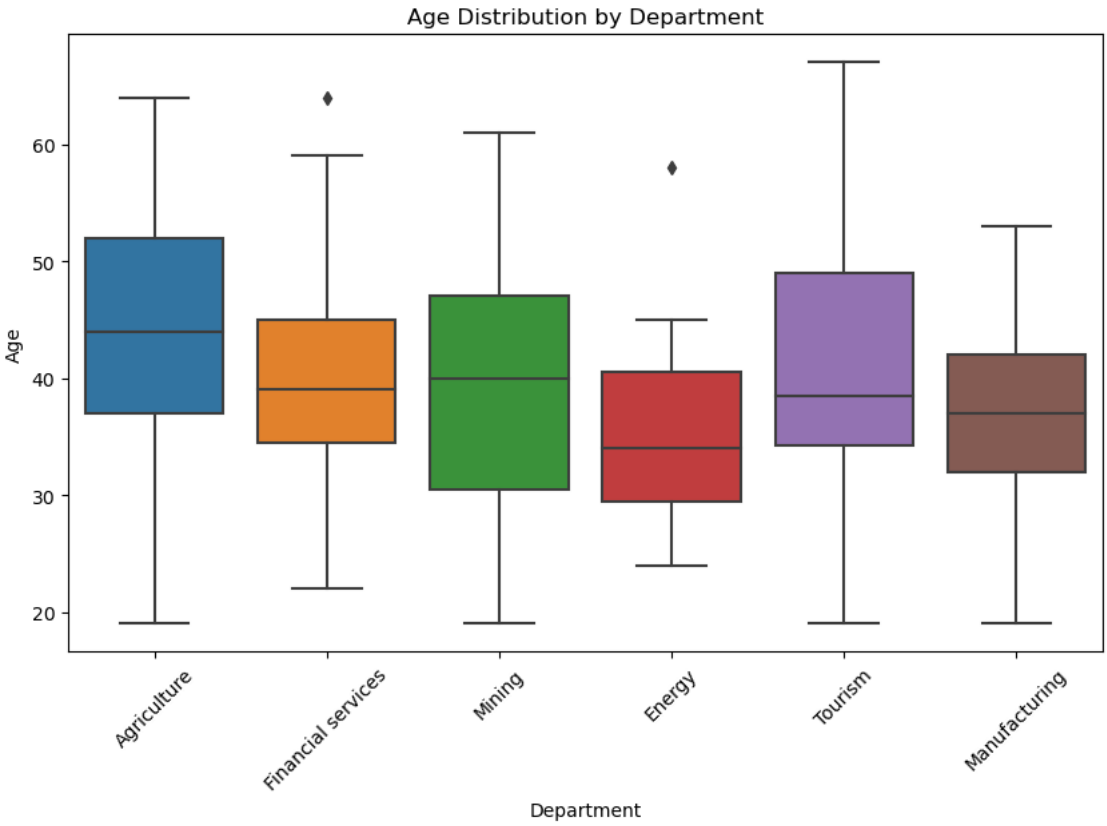
Now to perform analysis on the data: 1. Using appropriate graphs to represent the data

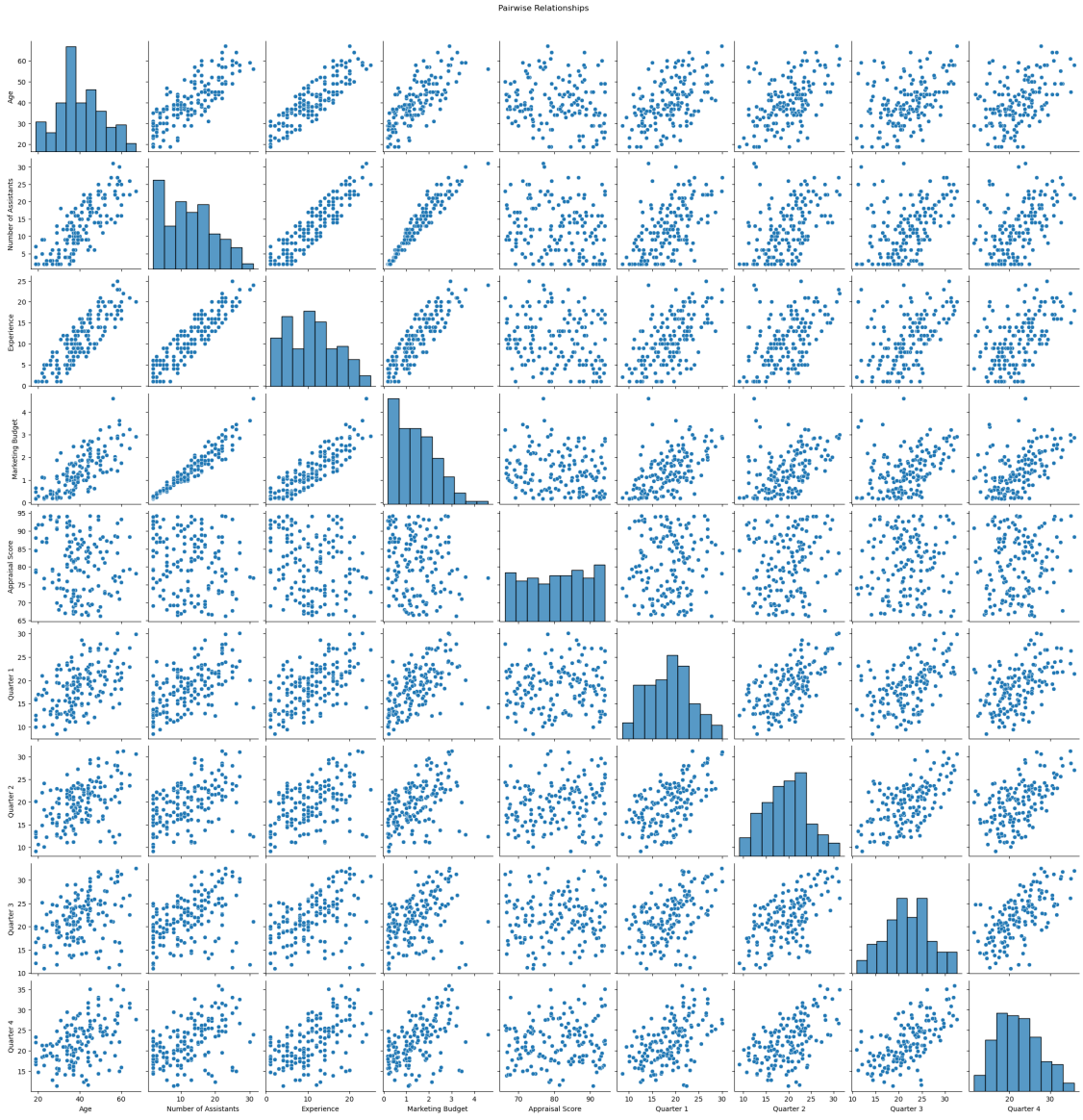
[4]:
```
#Bar plot for department-wise distribution
plt.figure(figsize=(10, 6))
sns.countplot(data=Sales_data, x='Department', palette='pastel')
plt.title('Distribution of Salespersons by Department')
plt.xlabel('Department')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

#Heatmap to visualize the correlations between the sales in each quarter and␣
 ↪other numerical variables usch as age, number of assistants, experience,␣
 ↪marketing budget and appraisal score
#Select the relevant numerical columns for the heatmap
numerical_columns = ['Age', 'Number of Assistants', 'Experience', 'Marketing␣
 ↪Budget', 'Appraisal Score', 'Quarter 1', 'Quarter 2', 'Quarter 3', 'Quarter␣
 ↪4']
#Calculate the correlation matrix
correlation_matrix = Sales_data[numerical_columns].corr()
#Create the heatmap using seaborn
plt.figure(figsize=(10, 8))
heatmap = sns.heatmap(correlation_matrix, annot=True, cmap="cividis")
plt.title('Correlation Heatmap')
plt.show()

#Box plot for age by department
plt.figure(figsize=(10, 6))
```

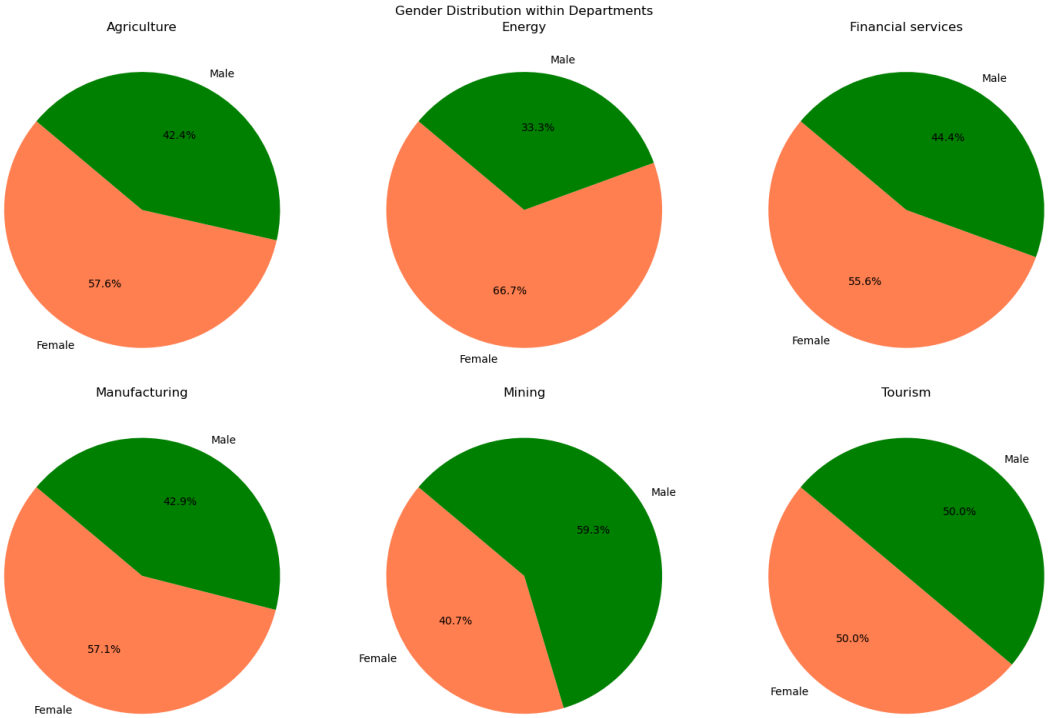
```
sns.boxplot(data=Sales_data, x='Department', y='Age')
plt.title('Age Distribution by Department')
plt.xlabel('Department')
plt.ylabel('Age')
plt.xticks(rotation=45)
plt.show()

#Pair plot for selected numerical columns
selected_numerical_columns = ['Age', 'Number of Assistants', 'Experience',␣
 ↪'Marketing Budget', 'Appraisal Score', 'Quarter 1', 'Quarter 2', 'Quarter␣
 ↪3', 'Quarter 4']
sns.pairplot(Sales_data[selected_numerical_columns])
plt.suptitle('Pairwise Relationships', y=1.02)
plt.show()

#Piecharts for gender distribution within each department
#Combined distribution of departments and gender
combined_distribution = Sales_data.groupby(['Department', 'Sex']).size().
 ↪unstack()
#Create a pie chart for each department
fig, axes = plt.subplots(2, 3, figsize=(15, 10))
fig.suptitle('Gender Distribution within Departments')
departments = combined_distribution.index
#Iterate through each department and create a pie chart
for i, (ax, department) in enumerate(zip(axes.flatten(), departments)):
    sizes = combined_distribution.loc[department].values
    ax.pie(sizes, labels=combined_distribution.columns, autopct='%1.1f%%',␣
 ↪startangle=140, colors=['coral','green'])
    ax.set_title(department)
plt.tight_layout()
plt.show()
```


Distribution of Salespersons by Department
Count
0
5
10
15
20
25
30
Agriculture
Financial services
Mining
Energy
Tourism
Manufacturing
Department


Correlation Heatmap
Age
Number of Assistants
Experience
Marketing Budget
Appraisal Score
Quarter 1
Quarter 2
Quarter 3
Quarter 4
1 0.79 0.87 0.76 -0.1 0.5 0.49 0.4 0.46
0.79 1 0.91 0.96 -0.13 0.57 0.48 0.4 0.5
0.87 0.91 1 0.88 -0.16 0.58 0.49 0.46 0.52
0.76 0.96 0.88 1 -0.12 0.56 0.45 0.38 0.49
-0.1 -0.13 -0.16 -0.12 1 0.075 0.16 0.039 0.12
0.5 0.57 0.58 0.56 0.075 1 0.59 0.48 0.52
0.49 0.48 0.49 0.45 0.16 0.59 1 0.6 0.56
0.4 0.4 0.46 0.38 0.039 0.48 0.6 1 0.65
0.46 0.5 0.52 0.49 0.12 0.52 0.56 0.65 1
1.0
0.8
0.6
0.4
0.2
0.0


Age Distribution by Department
Age
60
50
40
30
20
Agriculture
Financial services
Mining
Energy
Tourism
Manufacturing
Department


Pairwise Relationships
Age
Number of Assistants
Experience
Marketing Budget
Appraisal Score
Quarter 1
Quarter 2
Quarter 3
Quarter 4



PURPOSE OF THE GRAPHS PLOTTED ABOVE

Bar plot for Department Distribution: This graph provides a visual representation of the distribution of salespersons across different departments. It helps understand the composition of the sales force within each department and identify any differences in the number of salespersons across departments

Correlation heatmap: The heatmap provides a visual representation of the correlation between the numerical variables in the data set. It helps identify potential relationships and dependencies between different variables.

Box plot for age by department: This plot helps compare the distribution of ages across different departments, providing insights into the age profile within each department while identifying any variations or outliers.

Pair plot for selected numerical columns: The pair plot helps visualize the pairwise relationships between selected numerical variables, providing scatter plots for joint relationships and histograms for individual distributions.

Pie charts for gender distribution within each department: The visual representation offered by the pie charts serve the purpose of providing a focused view of the gender composition within each department at a glance

Each of these visualizations serve to convey specific information and insights such as distribution, relationships and patterns within the dataset. When used collectively, they provide a comprehensive understanding of the sales data and help derive actionable insights for decision making. NOTE:

There are so many ways to visualize the data but for illustration purposes I chose the few above.

2. Perform the relevant descriptive analysis

[5]:
```
#Calculate summary statistics for the numerical variables in the dataset
summary_statistics = Sales_data.describe()
print(summary_statistics)
```

```
              Age  Number of Assistants  Experience  Marketing Budget  \
count  153.000000            153.000000  153.000000        153.000000
mean    40.150327             12.385621   10.751634          1.426340
std     10.792768              7.277511    5.957394          0.889184
min     19.000000              2.000000    1.000000          0.170000
25%     34.000000              6.000000    6.000000          0.640000
50%     39.000000             12.000000   10.000000          1.290000
75%     47.000000             18.000000   15.000000          2.000000
max     67.000000             31.000000   25.000000          4.600000

       Appraisal Score   Quarter 1   Quarter 2   Quarter 3   Quarter 4
count       153.000000  153.000000  153.000000  153.000000  153.000000
mean         80.636601   18.563595   19.643856   21.874967   22.117059
std           8.355181    4.677800    4.807961    4.916638    5.292673
min          66.200000    8.570000    9.140000   10.950000   11.320000
25%          73.400000   15.000000   16.140000   18.770000   18.100000
50%          81.100000   18.800000   19.940000   21.900000   21.730000
75%          87.800000   21.780000   22.950000   24.950000   25.530000
max          94.200000   30.070000   31.310000   32.530000   35.910000
```

These statistics provide a comprehensive overview of the central tendency and dispersions for the numerical variables in the dataset,offering valuable insight into their characteristics.

Mean: Provides a measure of the center of the variable's distribution.

Std: The standard deviation reveals how much the value deviates from the mean, indicating the variable's spread and variability.

Range: The minimum and maximum values define the range of the variable, showcasing the minimum and maximum values within the dataset.

Quartiles: The quartiles help in understanding the spread and distribution of the dataset

3. Test the hypothesis of equality of proportions of female and male salespersons in each department and the entire organization

To test the hypothesis of equality of proportions of female and male salespersons in each department and the entire organization,we can perform a chi-squared test of independence. This test will help determine if there's an association between the gender of salespersons and their departments, as well as the organization overall. The null hypothesis(Ho) for each chi-squared test of independence would be that there is no association between gender and department/organization, while the alternative hypothesis(H1) would be that there is an association between gender and department.

[6]:
```
#Group the data by department and gender
sales_by_department_gender = Sales_data.groupby(['Department', 'Sex']).size().
 ↪unstack()
#Perform a chi-squared test for each department
department_tests = {}
for department in sales_by_department_gender.index:
    contingency_tables = sales_by_department_gender.loc[[department]]
    chi2, p, dof, expected = chi2_contingency(contingency_tables)
    department_tests[department] = {'chi2': chi2, 'p-value':p}
#Perform a chi-squared test for the entire organization
chi2, p, dof, expected = chi2_contingency(sales_by_department_gender.T)
organization_test = {'chi2': chi2, 'p-value':p}
department_tests, organization_test
```

[6]:
```
({'Agriculture': {'chi2': 0.0, 'p-value': 1.0},
  'Energy': {'chi2': 0.0, 'p-value': 1.0},
  'Financial services': {'chi2': 0.0, 'p-value': 1.0},
  'Manufacturing': {'chi2': 0.0, 'p-value': 1.0},
  'Mining': {'chi2': 0.0, 'p-value': 1.0},
  'Tourism': {'chi2': 0.0, 'p-value': 1.0}},
 {'chi2': 3.3384798815063323, 'p-value': 0.6479557790301915})
```

In each case the p-value is greater than the level of significance of 0.05, we therefore fail to reject the null hypothesis(Ho) at 5% level of significance and conclude that there is no association between the gender of a salesperson and their department as well as the organization as a whole

4. Is there a significant relationship between marital status and education level?

A common approach for this type of analysis is to perform a chi-squared test of independence.

[7]:
```
#Create a contingency table for the observed frequencies of marital status and␣
 ↪education level
contingency_table = pd.crosstab(Sales_data['Marital Status'],␣
 ↪Sales_data['Education Level'])
#Perform a chi-squared test of independence
chi2, p, dof, expected = chi2_contingency(contingency_table)
chi2, p
```

[7]:
```
(19.236160625508447, 0.023258092987928462)
```

With a p-value of 0.0233 we have evidence to reject the null hypothesis of independence between marital status and education level and conclude that there is a statistically significant relationship or association between these two categorical variables in the dataset. The results suggest that there is a non-random association between marital status and education level. The derived chi-squared value of 19.24 indicates a strong and significant association between marital status and education level suggesting that these two factors are not independent of each other within the dataset.

5. Is there a significant relationship between eduaction level and department?

[8]:
```
#Create a contingency table for the observed frequencies of education level and␣
 ↪department
contingency_table = pd.crosstab(Sales_data['Education Level'],␣
 ↪Sales_data['Department'])
#Perform a chi-squared test of independence
chi2, p, dof, expected = chi2_contingency(contingency_table)
chi2, p
```

[8]:
```
(20.61146622896623, 0.1497090924946819)
```

With a p-value of 0.1497, we don't have enough evidence to reject the null hypothesis of independence. Therefore, based on this analysis, there's no significant relationship between education level and department in the dataset.

6. Compare the mean quarterly and annual sales by age, sex, marital status, education level and department. For age categorize young employees as those aged less than 35 years and old employees otherwise

[9]:
```
#Categorize the employees into 'young' and 'old' based on the age variables
Sales_data['Age Group'] = np.where(Sales_data['Age'] < 35, 'Young', 'Old')
#Group the sales by specific demographic factors and calculate the mean sales
mean_sales_comparison = Sales_data.groupby(['Sex', 'Age Group', 'Marital␣
 ↪Status', 'Education Level', 'Department']).agg(
    {
        #Calculate both the mean and the total sum of sales for each quarter
        'Quarter 1' : ['mean', 'sum'],
        'Quarter 2' : ['mean', 'sum'],
        'Quarter 3' : ['mean', 'sum'],
        'Quarter 4' : ['mean', 'sum'],
    }
).reset_index()
#Calculate the mean annual sales
mean_sales_comparison['Annual'] = mean_sales_comparison[[('Quarter 1', 'mean'),␣
 ↪('Quarter 2', 'mean'), ('Quarter 3', 'mean'), ('Quarter 4', 'mean')]].
 ↪mean(axis=1)
#Replace the infinite and NaN values with np.nan
mean_sales_comparison.replace([np.inf, -np.inf], np.nan, inplace=True)
mean_sales_comparison.head()
```

[9]:

	Sex	Age Group	Marital Status	Education Level	Department	Quarter 1	\
						mean	
0	Female	Old	Divorced	Bachelors	Agriculture	21.605	
1	Female	Old	Divorced	Bachelors	Tourism	26.580	
2	Female	Old	Divorced	Certificate	Agriculture	18.200	
3	Female	Old	Divorced	Certificate	Mining	16.140	
4	Female	Old	Divorced	Diploma	Mining	21.110	

	Quarter 2	Quarter 3	Quarter 4	Annual

	sum	mean	sum	mean	sum	mean	sum	
0	43.21	23.965	47.93	26.67	53.34	27.595	55.19	24.95875
1	26.58	21.370	21.37	22.77	22.77	28.740	28.74	24.86500
2	18.20	23.110	23.11	24.47	24.47	30.920	30.92	24.17500
3	16.14	19.830	19.83	16.51	16.51	18.220	18.22	17.67500
4	42.22	22.940	45.88	25.65	51.30	24.105	48.21	23.45125

The interpretation of the results should consider the context of the organization and its market, as well as any specific goals or strategies that may benefit from the insights gained from comparison of mean sales figures across different demographic groups.

7. Is there a quarter where the sales were significantly different from others?

One approach to determine this is to perform an analysis of variance (ANOVA) test

[10]:
```
#Perform a one-way ANOVA to test for significant differences in mean sales␣
 ↪across quarters
f_statistic, p_value = f_oneway(Sales_data['Quarter 1'], Sales_data['Quarter␣
 ↪2'], Sales_data['Quarter 3'], Sales_data['Quarter 4'])
f_statistic, p_value
```

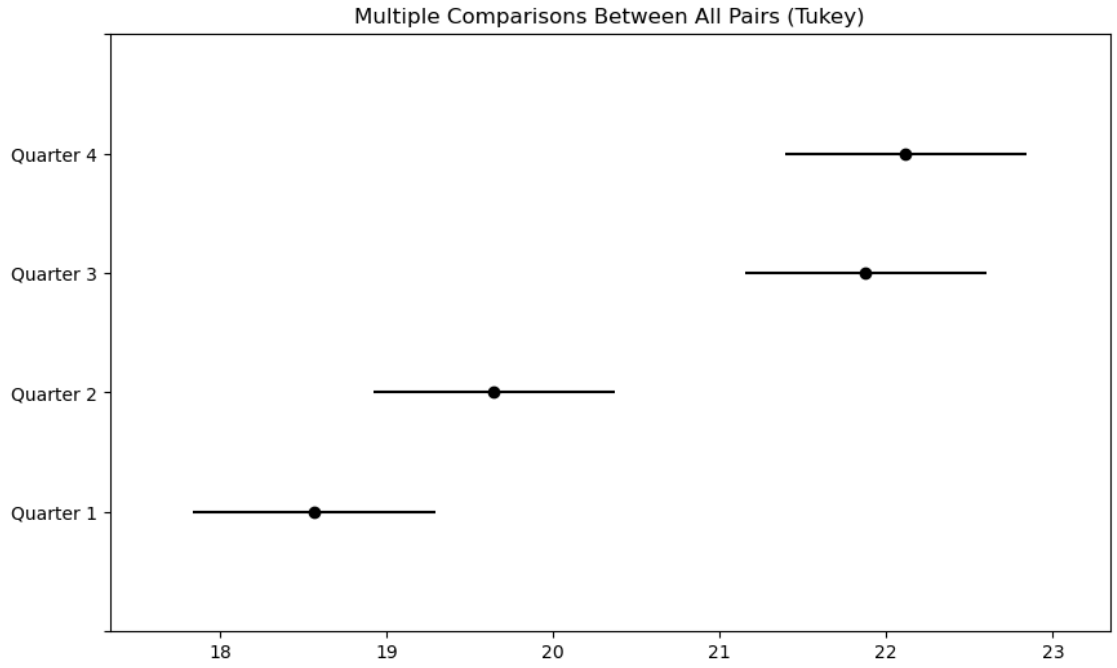
[10]:
```
(18.846054268034855, 1.065202213164001e-11)
```

The p-value (1.07e-11) is significantly less than a typical significance level (such as 0.05). Therefore, we have enough evidence to reject the null hypothesis and conclude that there are statistically significant differences in the mean sale figures across the four quarters. This suggests that at least one quarter has a significantly different mean sale figures compared to others and further test analyses can be conducted to determine which specific quarters differ significantly from each other in terms of sales performance. The post-hoc analysis to be conducted in this case is Tukey's range or pairwise t-test, which is a standard practice following a significant result from the ANOVA test.

[11]:
```
#Stack the quarterly sales data for the Tukey test
stacked_data = Sales_data[['Quarter 1', 'Quarter 2', 'Quarter 3', 'Quarter 4']].
 ↪stack()
stacked_data = pd.DataFrame({'sales': stacked_data, 'quarter': stacked_data.
 ↪index.get_level_values(1)})
#Perform Tukey's range test for post-hoc analysis
tukey_result = pairwise_tukeyhsd(stacked_data['sales'], stacked_data['quarter'])
#Generate the detailed results and summary table
tukey_plot = tukey_result.plot_simultaneous()
tukey_result.summary()
```

[11]:

group1	group2	meandiff	p-adj	lower	upper	reject
Quarter 1	Quarter 2	1.0803	0.222	-0.3715	2.5321	False
Quarter 1	Quarter 3	3.3114	0.0	1.8596	4.7632	True
Quarter 1	Quarter 4	3.5535	0.0	2.1017	5.0053	True
Quarter 2	Quarter 3	2.2311	0.0005	0.7793	3.6829	True
Quarter 2	Quarter 4	2.4732	0.0001	1.0214	3.925	True
Quarter 3	Quarter 4	0.2421	0.9734	-1.2097	1.6939	False



"group 1" and "group 2" are the compared quarters

"meandiff" presents the difference in mean sales figures between the quarters

"p-adj" provides the adjusted p-value for each comparison

8. Repeat 6) but use the non-parametric approach

For this particular approach, the Kruskal-Wallis H test can be employed

[12]:
```
#Categorizing employees into 'Young' and 'Old' based on the age variable
Sales_data['Age Group'] = Sales_data['Age'].apply(lambda age: 'Young' if age <␣
↪35 else 'Old')
#Performing the Kruskal-Wallis H test for non-parametric comparison
kruskal_result = kruskal(Sales_data['Quarter 1'], Sales_data['Quarter 2'],␣
↪Sales_data['Quarter 3'], Sales_data['Quarter 4'])
kruskal_result.statistic, kruskal_result.pvalue
```

[12]:
```
(48.059233596251794, 2.0685624593094217e-10)
```

The p-value(2.07e-10) is significantly less than a typical significance level(such as 0.05). Therefore, we have enough evidence to reject the null hypothesis and conclude that there are significant differences in the mean sales figures across the demographic groups. This suggests that atleast one demographic group has a significantly different mean sales figure compared to the others.

9. Repeat 7) but use the non-parametric approach

Yes, based on the results of the Kruskal-Wallis H test, which is a non-parametric approach, there is a strong evidence to suggest that there are significant differences in the mean sales figures across

the quarters. The low p-value obtained from the test indicates that at least one quarter has a significantly different mean sales figure compared to the others. Therefore, the non-parametric analysis suggests that there is indeed a quarter where the sales were significantly different from others.

10. Fit multiple linear regression models for quarterly and annual sales on age, number of assistants, experience, marketing budget and appraisal score. Comment on the adequecy of fit of the model and the significance of each of the independent variables

[13]:
```
#Define the independent variables for the regression models
independent_vars = ['Age', 'Number of Assistants', 'Experience', 'Marketing␣
 ↪Budget', 'Appraisal Score']
#Create a new column for the total quarterly sales
Sales_data['Total_Quarterly_Sales'] = Sales_data['Quarter 1'] +␣
 ↪Sales_data['Quarter 2'] + Sales_data['Quarter 3'] + Sales_data['Quarter 4']
#Fit the multiple linear regression models for the total quartserly sales
model_quarterly = sm.OLS(Sales_data['Total_Quarterly_Sales'], sm.
 ↪add_constant(Sales_data[independent_vars])).fit()
#Display the summary of the quarterly sales model
print(model_quarterly.summary())
```

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:     Total_Quarterly_Sales   R-squared:
0.441
Model:                               OLS   Adj. R-squared:
0.422
Method:                    Least Squares   F-statistic:
23.15
Date:                   Tue, 28 Nov 2023   Prob (F-statistic):
4.19e-17
Time:                           12:51:04   Log-Likelihood:
-598.13
No. Observations:                    153   AIC:
1208.
Df Residuals:                        147   BIC:
1226.
Df Model:                              5
Covariance Type:               nonrobust
==============================================================================
========
                             coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const                     26.2620     10.908      2.408      0.017       4.706
47.818
```

```
Age                      0.0893      0.191      0.468      0.640      -0.288
0.466
Number of Assistants     0.5140      0.538      0.956      0.341      -0.549
1.577
Experience               1.2429      0.504      2.467      0.015       0.247
2.239
Marketing Budget        -1.3033      3.882     -0.336      0.738      -8.975
6.368
Appraisal Score          0.4276      0.122      3.517      0.001       0.187
0.668
==============================================================================
Omnibus:                       12.494   Durbin-Watson:                   1.677
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               13.093
Skew:                          -0.656   Prob(JB):                      0.00143
Kurtosis:                       3.577   Cond. No.                     1.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R-squared and Adjusted R-squared: The R-squared value(0.441) indicates that approximately 44.1% of the variation in the total quarterly sales can be explained by the independent variables in the model. The adjusted R-squared, which accounts for the number of predictors in the model, suggests the same.

F-statistic: The F-statistic tests the overall significance of the regression model. In this case, the F-statistic is statistically significant , indicating the regression model as a whole provides a better fit than a model with no independent variables.

Significance of Independent Variables: Among the independent variables, experience and appraisal score exhibit statistical significance, with p-values < 0.05. This suggests that these variables are statistically significant in explaining the variation in the total quarterly sales. In contrast, the coefficients for age, number of assistants and marketing budget do not appear to be statistically significant, as indicated by their p-values(>0.05)

Condition Number: The high condition number raises the potential issue of multicollinearity, suggesting that there may be some degree of correlation between the independent variables.

In summary, while the model explains a moderate proportion of the variations in total quarterly sales and some of the independent variables appear to be statistically significant, the large condition number suggests potential issues with multicollinearity. This may require further investigation and potentially refinement of the model to address. It's important to interpret these findings in the context of the specific goals and requirements of the analysis, as well as to consider the practical significance of the results within the relevant domains.

[ ]: