Quantium Virtual Internship - Retail Strategy and Analytics

Task:

Our client has asked us to evaluate the performance of a store trial which was performed in stores 77, 86 and 88.

This can be broken down by:

total sales revenue total number of customers average number of transactions per customer

Create a measure to compare different control stores to each of the trial stores to do this write a function to reduce having to re-do the analysis for each trial store. Consider using Pearson correlations or a metric such as a magnitude distance e.g. 1- (Observed distance – minimum distance)/(Maximum distance – minimum distance) as a measure.

Once you have selected your control stores, compare each trial and control pair during the trial period. You want to test if total sales are significantly different in the trial period and if so, check if the driver of change is more purchasing customers or more purchases per customers etc.

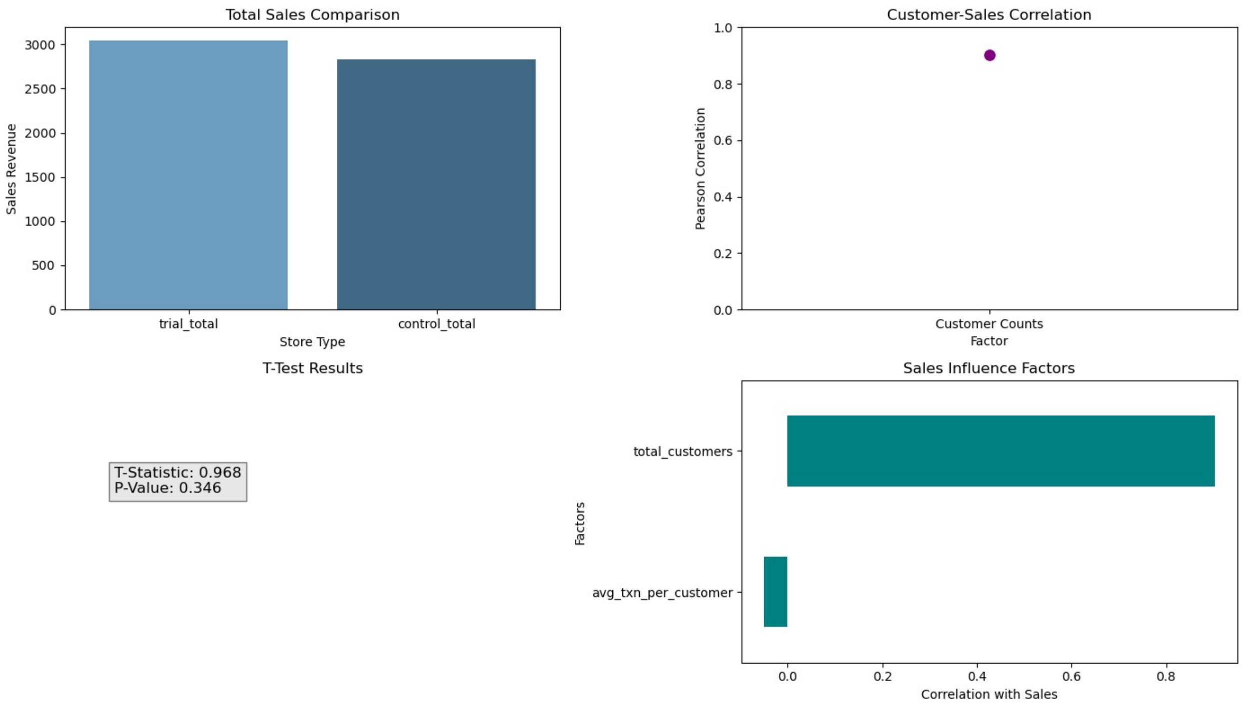
Solution:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr, ttest_ind

# Load and prepare the dataset
file_path = 'QVI_data.xlsx'
qvi_data = pd.read_excel(file_path)

# Data Cleaning and Preparation
qvi_data = qvi_data.dropna(subset=['STORE_NBR', 'TOT_SALES', 'DATE'])
qvi_data['STORE_NBR'] = qvi_data['STORE_NBR'].astype(int)
qvi_data['MONTH'] = qvi_data['DATE'].dt.to_period('M')

# Filter trial and control stores
trial_stores = [77, 86, 88]
control_stores = [233, 89, 168]
filtered_data = qvi_data[qvi_data['STORE_NBR'].isin(trial_stores +
control_stores)]

# Calculate monthly metrics
monthly_metrics = filtered_data.groupby(['STORE_NBR', 'MONTH']).agg(
    total_sales=('TOT_SALES', 'sum'),
    total_customers=('LYLTY_CARD_NBR', 'nunique'),
```

```
    total_transactions=('TXN_ID', 'count')
).reset_index()

monthly_metrics['avg_txn_per_customer'] = (
    monthly_metrics['total_transactions'] /
monthly_metrics['total_customers']
)

# Analysis Function
def analyze_stores(trial_store, control_store, metrics_df):
    trial_data = metrics_df[metrics_df['STORE_NBR'] == trial_store]
    control_data = metrics_df[metrics_df['STORE_NBR'] ==
control_store]

    common_months =
set(trial_data['MONTH']).intersection(control_data['MONTH'])
    trial_data = trial_data[trial_data['MONTH'].isin(common_months)]
    control_data =
control_data[control_data['MONTH'].isin(common_months)]

    trial_sales = trial_data['total_sales']
    control_sales = control_data['total_sales']

    customer_corr, _ = pearsonr(trial_data['total_customers'],
trial_data['total_sales'])
    t_stat, p_value = ttest_ind(trial_sales, control_sales,
equal_var=False)
    sales_correlation = trial_data[['total_customers',
'avg_txn_per_customer']].corrwith(trial_data['total_sales'])

    return {
        'sales_comparison': {'trial_total': trial_sales.sum(),
'control_total': control_sales.sum()},
        'customer_sales_correlation': customer_corr,
        't_test': {'t_stat': t_stat, 'p_value': p_value},
        'sales_influence_factors': sales_correlation,
    }

# Example Analysis
results = analyze_stores(77, 233, monthly_metrics)

# Visualization
plt.figure(figsize=(14, 8))

# Subplot 1: Sales Comparison
plt.subplot(2, 2, 1)
sns.barplot(
    x=list(results['sales_comparison'].keys()),
    y=list(results['sales_comparison'].values()),
    palette='Blues_d',
```

```python
    hue=None
)
plt.title('Total Sales Comparison')
plt.ylabel('Sales Revenue')
plt.xlabel('Store Type')

# Subplot 2: Customer-Sales Correlation
plt.subplot(2, 2, 2)
sns.scatterplot(x=['Customer Counts'],
y=[results['customer_sales_correlation']], s=100, color='purple')
plt.title('Customer-Sales Correlation')
plt.ylim(0, 1)
plt.ylabel('Pearson Correlation')
plt.xlabel('Factor')

# Subplot 3: T-Test Results
plt.subplot(2, 2, 3)
plt.text(
    0.1, 0.6,
    f"T-Statistic: {results['t_test']['t_stat']:.3f}\nP-Value:
{results['t_test']['p_value']:.3f}",
    fontsize=12,
    bbox=dict(facecolor='lightgrey', alpha=0.5)
)
plt.title('T-Test Results')
plt.axis('off')

# Subplot 4: Sales Influence Factors
plt.subplot(2, 2, 4)
results['sales_influence_factors'].sort_values().plot(kind='barh',
color='teal')
plt.title('Sales Influence Factors')
plt.xlabel('Correlation with Sales')
plt.ylabel('Factors')

plt.tight_layout()
plt.show()
```

```
C:\Users\robin\AppData\Local\Temp\ipykernel_14816\2680376621.py:62:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(
```



Results and Findings:

1. Total Sales Comparison: The bar plot compares total sales revenue between the trial store (trial_total) and the control store (control_total).

   ## Observation:

   The trial store's total sales (~3040) are slightly higher than the control store's total sales (~2827). This suggests that the trial store might have experienced a sales boost, but further analysis (like statistical significance) is needed to confirm.

2. Customer-Sales Correlation: This scatter plot displays the Pearson correlation coefficient (0.903) between customer counts and total sales for the trial store.

   ## Observation:

   A high positive correlation (close to 1) suggests that increasing customer counts significantly drives sales revenue for the trial store.

3. T-Test Results: This text box displays the results of a T-Test comparing sales between the trial and control stores: T-Statistic: 0.968 P-Value: 0.346

   ## Observation:

   The p-value (0.346) is much greater than the typical significance threshold (0.05), indicating that the difference in sales between the trial and control stores is not statistically significant. Therefore, the observed difference might be due to chance.

4. Sales Influence Factors: A horizontal bar plot shows the correlation of two factors with sales for the trial store: total_customers: 0.903 (strong positive correlation) avg_txn_per_customer: -0.05 (weak negative correlation)

Observation:

Sales are strongly influenced by the number of customers, whereas the average number of transactions per customer has a negligible and slightly negative effect.

Summary of Insights:

Trial Store Performance:

Sales are slightly higher in the trial store compared to the control store, but the difference is not statistically significant.

Key Driver of Sales:

The increase in customer counts is the primary factor influencing sales performance. Efforts to drive sales should focus on attracting more customers rather than increasing transactions per customer.

Statistical Significance:

The lack of statistical significance suggests that the observed differences might not be attributable to the trial intervention. This analysis provides actionable insights for strategizing sales improvement initiatives.