Module 2 Data Science Project: Orders

David Goldstein

dgoldstein24@gmail.com
dg2996@columbia.edu

Obtain and Scrub: Purchase Order Data Using SQL and change format to DataFrame

```
#connect sqlite3 to the Northwind Database
connection = sqlite3.connect('Northwind_small.sqlite')
```

```
#Attach cursor to connection in order to execute SQL commands
cursor = connection.cursor()
```

```
#Execute SQL command to obtain OrderID table information
order_details = cursor.execute('''SELECT * FROM OrderDetail;''').fetchall()

#Convert query into Pandas DataFrame
details_df = pd.DataFrame(order_details, columns = ['x','OrderID',
                                                    'ProductID', 'UnitPrice',
                                                    'Quantity', 'Discount'])

#drop irrelevant columns
details_df = details_df.drop(['x'], axis = 1)

#show DataFrame
details_df.head()
```

	OrderID	ProductID	UnitPrice	Quantity	Discount
0	10248	11	14.0	12	0.0
1	10248	42	9.8	10	0.0
2	10248	72	34.8	5	0.0
3	10249	14	18.6	9	0.0
4	10249	51	42.4	40	0.0

Hypotheses: Prompt

Null hypothesis: there is no significant difference between order quantity with or without a discount. In other words, the mean difference between the discount and non-discount group is zero

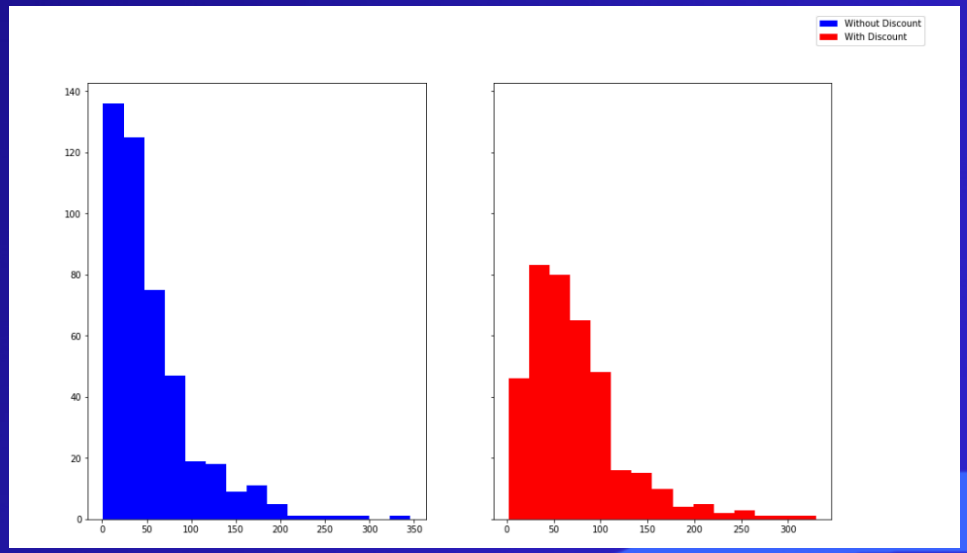
Alternative hypothesis: a discount significantly affects the quantity of items purchased in an order

This test is two-tailed: although a negative relationship is unlikely, the question asks for an overall effect made by disount, not just an increase

Explore (Prompt Pt 1): Histogram visualizations and conditions for student t-test

Notes:

- Lacks normality
- Skewed right
- Different Sample Sizes
- ***Conclusion: Use General T-Test***



Model (Prompt Pt 1): Calculate p-value using general test

```
#Find T statistic and pvalue using scipy.stats
stats.ttest_ind(with_discount['Quantity'],
                without_discount['Quantity'], equal_var = False)
```

T-STAT: 5.89

P-VALUE: 5.85 e – 09

P < .05 ---→ result is significant

But does the finding have a large effect at this level of confidence?

Model: Calculate Cohen D – measure for effect size

```
#Test for effect size using Cohen'd d that measures
#effect size of samples with differing means

def cohend(sample_1, sample_2):
    n1 = len(sample_1)
    n2 = len(sample_2)
    s1 = np.var(sample_1, ddof = 1)
    s2 = np.var(sample_2, ddof = 1)
    u1 = sample_1.mean()
    u2 = sample_2.mean()
    pooled_standard_deviation = np.sqrt((((n1-1)*s1)+((n2-1)*s2))/(n1+n2-2))
    #weighted std for the two samples together

    cd = (u1-u2) / pooled_standard_deviation
    return cd
```

Cohen D = .421

There is a weak/moderate positive relationship between discounts and quantity

Prompt Pt 2: Significance & Effect Size of Varying Discount Levels

Step 1: Create DataFrame for each Discount Level

Step 2: Loop to find p-value & effect size for each Discount Level

Business Conclusion*: Discounts increase quantities across the board. 25% discount has the most reliable and biggest positive impact on quantity**

Discount	P-Value	Effect Size
• .05	.00199	.365
• .10	.00089	.415
• .15	.00085	.486
• .20	.01037	.338
• .25	.00069	.505

Next Step: Edit DataFrame to include columns for further analysis

Loop Through SQL Statements

```python
#Create lists that will be turned into columns in df_2
corresponding_regions = []
corresponding_title = []

#Create loop to fill lists
for i in range(len(df_2)):

    #format each customer code to work with SQL command syntax
    code = "'" + str(df_2.iloc[i][0]) + "'"

    #add customer code to commands to fetch region
    region_command = 'SELECT Region FROM customers WHERE Code = ' + code

    #add customer code to commands to fetch title
    title_command = 'SELECT ContactTitle FROM customers WHERE Code = ' + code

    #execute command to fetch region
    region = cursor.execute(region_command).fetchall()

    #execute command to fetch region
    title =  cursor.execute(title_command).fetchall()

    #add region and title to corresponding list
    corresponding_regions.append(region)
    corresponding_title.append(title)

#add regions and titles as coulumns in df_2
df_2['Region'] = corresponding_regions
df_2['Title'] = corresponding_title
```

Scrub text data

Region	Title
[(Western Europe,)]	[(Accounting Manager,)]
[(Western Europe,)]	[(Marketing Manager,)]
[(South America,)]	[(Accounting Manager,)]
[(Western Europe,)]	[(Sales Agent,)]
[(Western Europe,)]	[(Accounting Manager,)]

Region	Title
Western Europe	Accounting Manager
Western Europe	Marketing Manager
South America	Accounting Manager
Western Europe	Sales Agent
Western Europe	Accounting Manager

Updated DataFrame

	CustomerID	Freight	OrderID	Quantity	Discount	Region	Title
0	VINET	32.38	10248	27	0.00	Western Europe	Accounting Manager
1	TOMSP	11.61	10249	49	0.00	Western Europe	Marketing Manager
2	HANAR	65.83	10250	60	0.15	South America	Accounting Manager
3	VICTE	41.34	10251	41	0.05	Western Europe	Sales Agent
4	SUPRD	51.30	10252	105	0.05	Western Europe	Accounting Manager

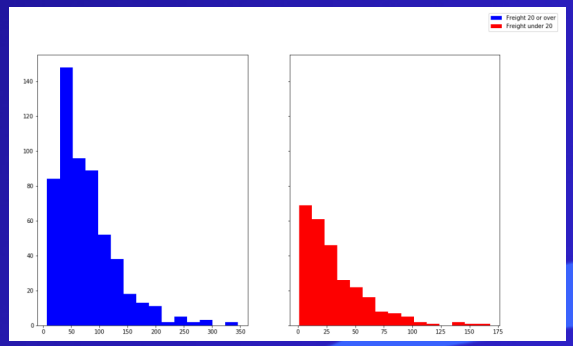
Original question 1: Do higher shipping costs affect purchase quantity?

Null Hypothesis 1: A shipping charge of 20 or more does not signficantly affect quantity of items purchased

Alternative Hypothesis: A shipping charge of 20 or more signficantly affects quantity of items purchased

Notes:

- Lacks normality
- Skewed right
- Different Sample Sizes
- ***Conclusion: Use General Test***



Question 1: T-Stat, P-Value, Effect Size, and Interpretation

P-VALUE: 5.85 e – 09

P < .05 ---→ result is significant

Effect Size = .941 ---→ positive, very strong relationship

Conclusion: The business should look into offering free shipping as a means to increase quantity purchased. This strategy mimics Amazon Prime's free shipping strategy

Original question 2:
Do companies from Western Europe order more?

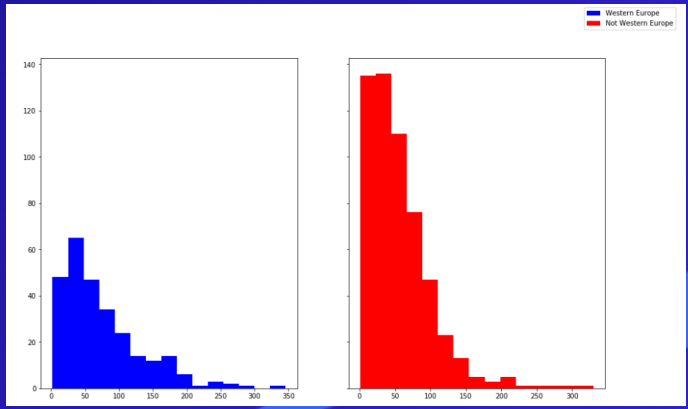
Null Hypothesis 2: A purchase from Western Europe does not signficantly affect quantity of items purchased

Alternative Hypothesis: A purchase from Western Europe signficantly affects quantity of items purchased

Notes:

- Lacks normality
- Skewed right
- Different Sample Sizes
- ***Conclusion: Can't use normal t-test***

USE WELCH'S T-TEST



Question 2: T-Stat, P-Value, Effect Size, and Interpretation

P-VALUE: 5.26 e – 06

P < .05 ---→ result is significant

Effect Size = .372---→ positive, weak relationship

Conclusion: The business should look consider marketing its products in Western Europe, as there is statistical backing to believe that customers from there purchase slightly greater quantities of goods than the rest of the world

Original question 3:
Do owners order more or less than other employees?

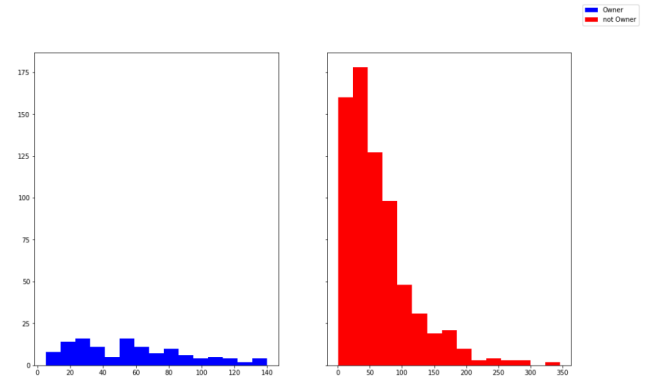
Null Hypothesis 3: Being an owner calling for the order does not signficantly affect quantity of items purchased

Alternative Hypothesis: Being an owner calling for the order signficiantly affects quantity of items purchased

Notes:

- Not owner is skewed right
- Owner graph might be very slightly normal
- Different Sample Sizes
- ***Conclusion: Can't use normal t-test***

USE WELCH'S T-TEST



Question 3: T-Stat, P-Value, Effect Size, and Interpretation

P-VALUE: .09

P > .05 ---→ result is INSIGNIFICANT

Effect Size = -.124 --→ weak, negative relationship

Conclusion: The business cannot have confidence in the conclusion that owners purchase more/fewer goods than their employees. Therefore, the business should not expend resources marketing its products to owners as opposed to oother types of customers

Interpret: Recommendations

1. Rec 1: Utilize discounts, especially 25%, to boost sale quantity that are lagging in purchases. ✓
2. Rec 2: Offer free shipping as a means of boosting purchase quantity ✓
3. Rec 3: Focus marketing on Western Europe ✓
4. Rec 4: Do not spend marketing budget on targeted appeals to business owners ✗

Interpret: Project Takeaways

1. Combining SQL and Pandas is a powerful way to obtain and sort data

2. Checking for necessary conditions for t-test is essential

3. Knowing p-value is usually not enough – it should be evaluated alongside effect size

Thank You

David Goldstein