

MI-462 Final Project

Daniel DeLuca, Colin Rehnert, Matt Granato

Outline

- Project Overview
- Variables in the Dataset
- Microsoft Azure Context
- Power BI Context
- Comparison of Azure and Power BI
- Question 1 in Azure and Power BI
- Question 2 in Azure and Power BI
- Question 3 in Azure and Power BI
- Question 4 in Azure and Power BI
- Question 5 in Azure and Power BI
- Question 6 on Data Ethics
- Conclusion

Project Overview

- For our project, we decided to work on a dataset that involves shipping goods with different modes of transportation.
- The dataset is called Train.csv
- We are using Azure Data Explorer to make SQL/KQL Queries and Visualizations to answer our questions.
- We are also using PowerBI to create visualizations with Python script to help make our data more understandable.

Variables in Our Dataset

ID

Warehouse_block (A,B,C,D,E,F)

Mode_of_Shipment (Flight, Road, Ship)

Customer_care_calls (2,3,4,5,6,7)

Customer_rating (1,2,3,4,5)

Cost_of_the_Product (range)

Prior_purchases (range)

Product_importance (low, medium, high)

Gender (M,F)

Discount_offered (range)

Weight_in_gms (range)

Reached.on.Time_Y.N (1= 'Yes', 0='No')

Microsoft Azure

Our team decided to use Microsoft Azure to analyze and create visualizations for our dataset. We started our process in Azure by creating an Azure data explorer cluster. We then linked the URI of this cluster with Azure Data Explorer. The next step was to ingest our data set so we could start to perform SQL/KQL queries on it. Once we ingest our data we formulated a list of questions about our dataset we wanted to answer. We then answered these questions using Chat GPT to write our code.

Power BI

After creating some visualizations and working with the data in Azure, we realized that we wanted to explore using another program to compare the two and learn new skills. We decided that Power BI would be the best option as it uses Python code instead of SQL/KQL. We downloaded Power BI to our individual computers and we able to upload our data set through Power BI. Following that process we built our the visualizations by dropping in the columns that we wanted to compare.

Comparison of Code Generation with Power BI (Python) and Azure (SQL/KQL)

1. Create a Stacked Bar Chart:

- In the 'Report' view, drag 'Warehouse_block' to the Axis and 'Reached.on.Time_Y.N' to the Values of a stacked bar chart.
- This will display the count of on-time deliveries for each warehouse block.

2. Sort the Stacked Bar Chart:

- Sort the chart by 'Reached.on.Time_Y.N' in descending order to easily identify the warehouse block with the highest count of on-time deliveries.

3. Create a Card Visualization for the Best Block:

- Create a card visualization.
- Use the following DAX formula to find the warehouse block with the highest count of on-time deliveries:

```
DAX
BestBlock =
CALCULATE(
    VALUES('Train'[Warehouse_block]),
    TOPN(1, ALL('Train'), 'Reached.on.Time_Y.N', DESC)
)
```

- Display the 'BestBlock' measure in the card visualization.

DA

there are two columns in dataset, one column is Warehouse_block and the possible options for this column are A,B,C,D,E,F. the other column is Reached.on.Time_Y.N with the options of 1,0. Write KQL code that shows the percentage of 1 and 0 for each warehouse block compared to the total for each warehouse block

kql

Copy code

```
let datasetSummary = dataset
| summarize TotalCount = count() by Warehouse_block
| extend Count1 = countif(['Reached.on.Time_Y.N'] == 1),
    Count0 = countif(['Reached.on.Time_Y.N'] == 0);
datasetSummary
| project Warehouse_block, Count1, Count0,
    Percentage1 = round(100.0 * todouble(Count1) / todouble(TotalCount), 2),
    Percentage0 = round(100.0 * todouble(Count0) / todouble(TotalCount), 2)
```

Chat GPT was able to easily generate Python Code and give detailed directions for how to use Power BI.

Chat GPT would struggle to create code compatible with Azure. It needed a detailed description of the data and would usually need to be modified several times before it was compatible with Azure.

Question 1: What is the gender distribution and are there any differences in customer rating between males and females? (Azure)

Conclusion:

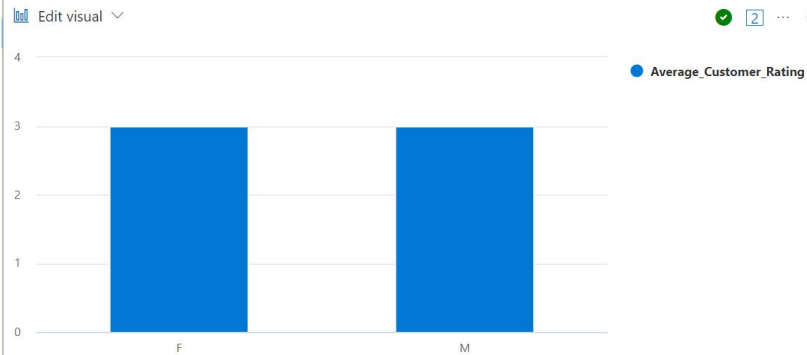
The gender distribution is 5545 females and 5454 males. There seems to be almost no difference in customer rating based on gender.

```
1 let datasetSummary = dataset
2   | summarize Count = count() by Gender;
3 datasetSummary
4   | project Gender, Count
5
```

Table 1 + Add visual Stats Search

	Gender	Count
>	F	5,545
>	M	5,454

```
1 let datasetSummary = dataset
2   | summarize Average_Customer_Rating = avg(Customer_rating) by Gender;
3 datasetSummary
4   | project Gender, Average_Customer_Rating
```



```
1 let datasetSummary = dataset
2   | summarize Average_Customer_Rating = avg(Customer_rating) by Gender;
3 datasetSummary
4   | project Gender, Average_Customer_Rating
```

Table 1 + Add visual Stats Search UTC Done (0.154 s) 2

	Gender	Average_Customer_Rating
>	F	2.986654643823264
>	M	2.9944994499449944

Question 1: What is the gender distribution and are there any differences in customer rating between males and females? (Power BI)

Count of Gender:

Female: 5,545

Female %: 50.41%

Male: 5,454

Male %: 49.59%

Avg. Customer
Rating:

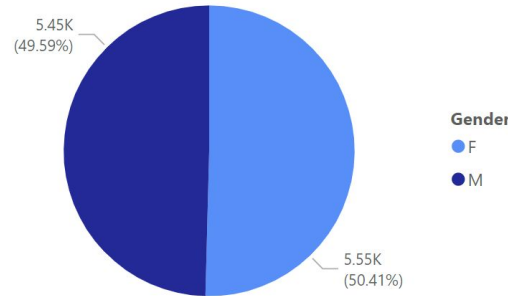
Female avg: 2.99

Male avg: 2.99

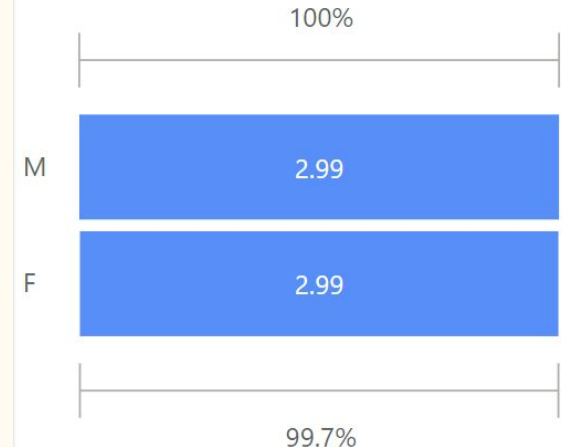
Conclusions:

- Slightly larger population of females over males
- Avg. customer rating was not different between males and females

Count of Gender by Gender



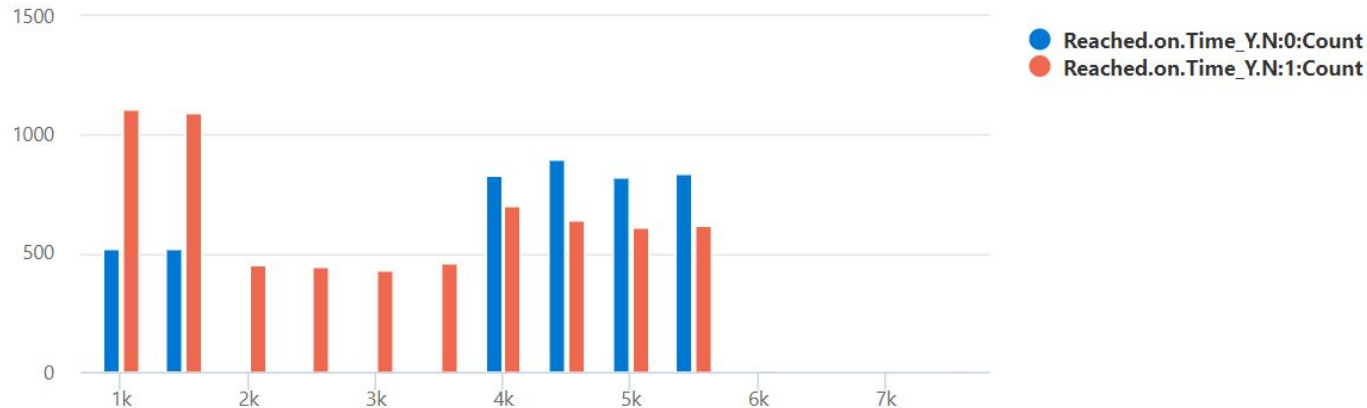
Average Customer Rating by Gender



Question 2: Does the weight of a product affect on time status? (Azure)

```
1 SELECT
2     (FLOOR([Weight_in_gms] / 500) * 500) AS Weight_Range,
3     [Reached.on.Time_Y.N],
4     COUNT(*) AS Count
5 FROM dataset
6 GROUP BY (FLOOR([Weight_in_gms] / 500) * 500), [Reached.on.Time_Y.N]
7 ORDER BY Weight_Range, [Reached.on.Time_Y.N];
```

Edit visual



Conclusion:
As the weight of the product increased, the chance of the product not being on time increased.

Question 2: Does the weight of a product affect on time status? (Power BI)

Weights v. Reached on Time:

On time: 1

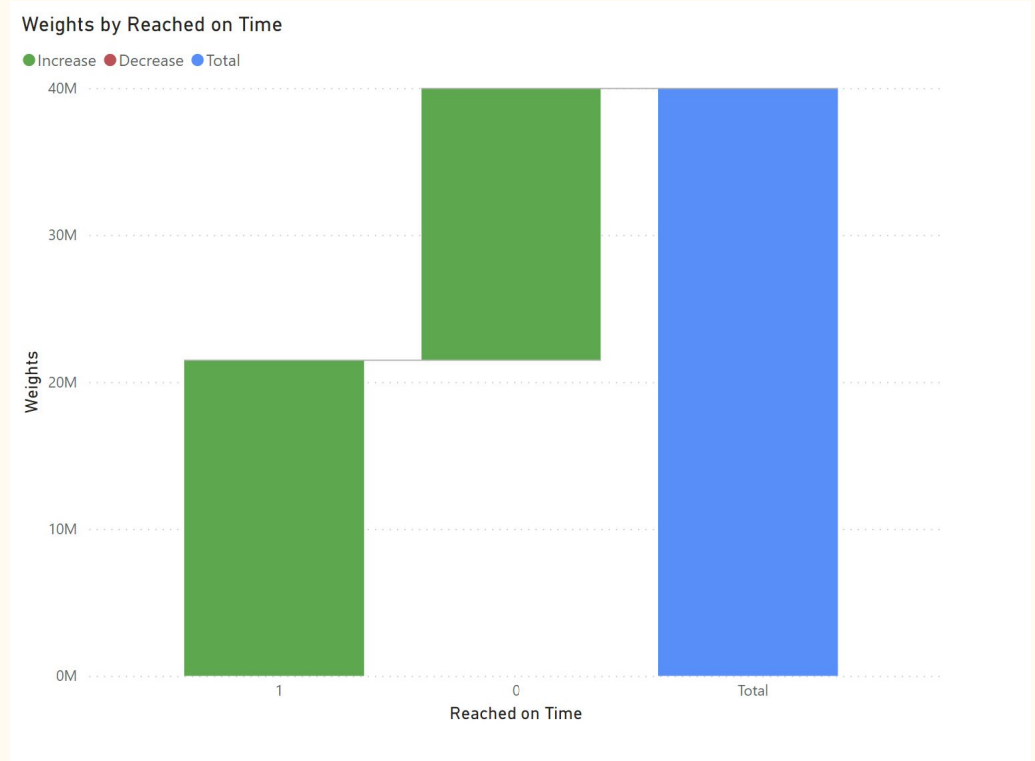
Late: 0

Weight:

0-40 million grams

Conclusion:

- The shipments that weigh more than 21 million grams tend to be later than those less than 21 million grams



Question 3: What percentage of deliveries are on time? What is the most common mode of transportation? (Azure)

Conclusion: There are 6,563 on time deliveries or 59.67% of total deliveries, and 4,436 not on time deliveries or 40.33% of total deliveries.

```
1 SELECT
2     [Reached.on.Time_Y.N],
3     COUNT(*) AS Count
4 FROM dataset
5 GROUP BY [Reached.on.Time_Y.N];
```

Reached.on.Time_Y.N	Count
1	6,563
0	4,436

```
1 dataset
2 | summarize Percentage_0 = round(100.0 * countif(['Reached.on.Time_Y.N'] == 0) / count(), 2),
3 |           Percentage_1 = round(100.0 * countif(['Reached.on.Time_Y.N'] == 1) / count(), 2)
4 | project Percentage_0, Percentage_1
5
```

Percentage_0	Percentage_1
40.33	59.67

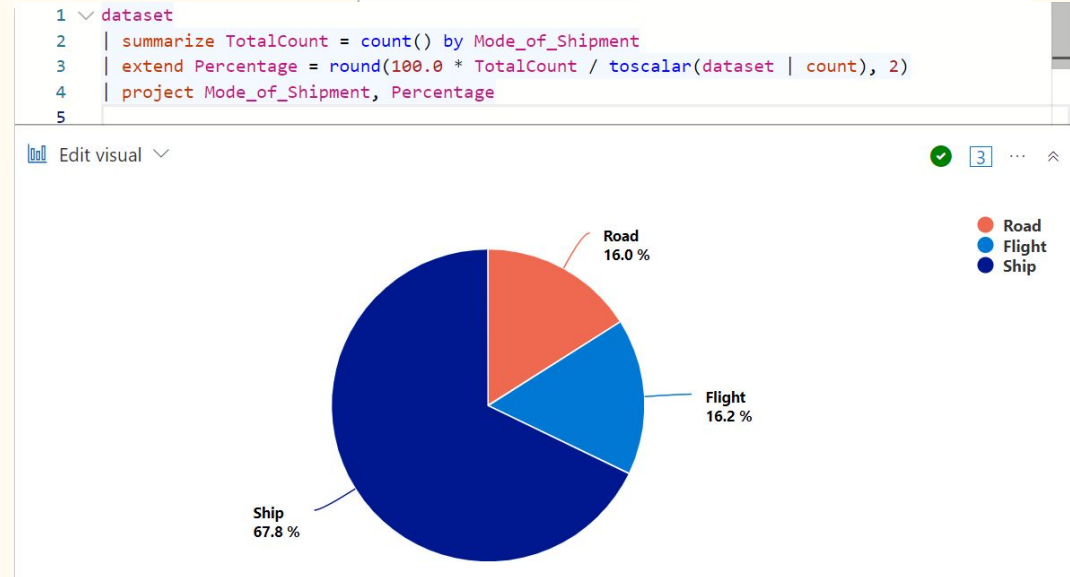
Question 3 (2): What percentage of deliveries are on time? What is the most common mode of transportation? (Azure)

Conclusion: The most common type of delivery is Ship with 67.8% followed by Flight with 16.2% and then least common delivery type Road at 16%.

```
1 dataset
2 | summarize TotalCount = count() by Mode_of_Shipment
3 | extend Percentage = round(100.0 * TotalCount / toscalar(dataset | count), 2)
4 | project Mode_of_Shipment, Percentage
5
```

Table 1 + Add visual © Stats 🔍 Search ⌚ UTC ✅ Done (0.049 s) 📄 3 records

Mode_of_Shipment	Percentage
> Ship	67.84
> Flight	16.16
> Road	16



Question 3: What percentage of deliveries are on time? What is the most common mode of transportation? (Power BI)

Deliveries Y/N: (Y=1), (N=0) Conclusions:

Y %: 59.67%

N %: 40.33%

- Shipping is the most dominant form of transportation despite being the slowest
- The gap between deliveries being on schedule or not on schedule is smaller than we originally expected when looking at the data

Modes of Transport:

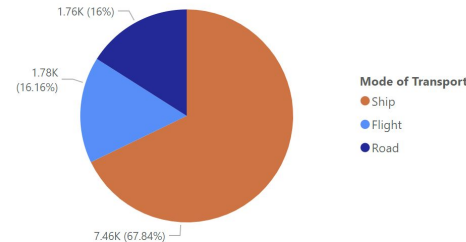
(In order of usage)

Ship: 67.84%

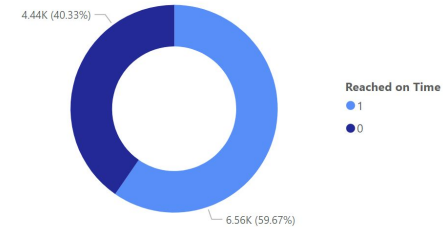
Flight: 16.16%

Road: 16%

Mode Usage by Mode of Transport



Yes or No by Reached on Time



Question 4: Which is each warehouse block's percentage of on time delivery, which warehouse block is the best? (Azure)

```
1 let datasetSummary = dataset
2   | summarize Count1 = countif(['Reached.on.Time_Y.N'] == 1),
3     Count0 = countif(['Reached.on.Time_Y.N'] == 0)
4   by Warehouse_block;
5 datasetSummary
6 | extend TotalCount = Count1 + Count0
7 | extend Percentage1 = round(100.0 * todouble(Count1) / todouble(TotalCount), 2),
8   Percentage0 = round(100.0 * todouble(Count0) / todouble(TotalCount), 2)
9 | project Warehouse_block, Count1, Count0, Percentage1, Percentage0
10
```

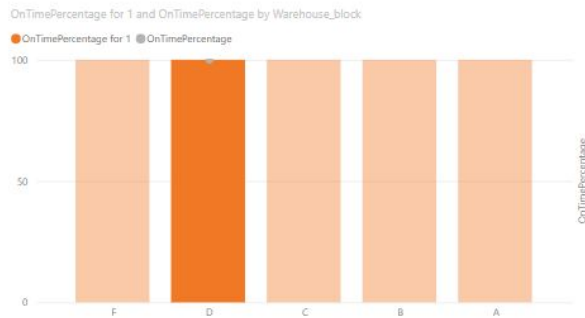
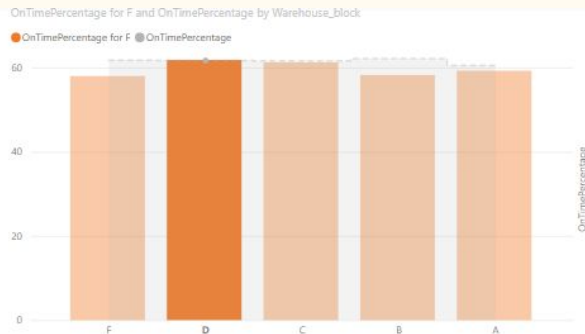
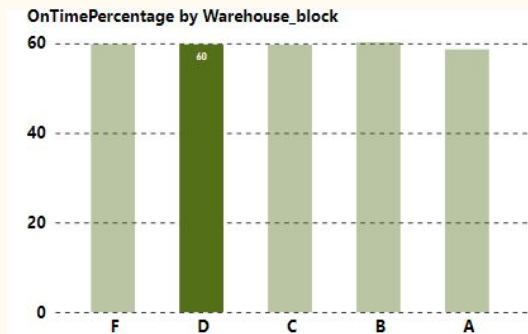
Table 1 + Add visual © Stats 🔍 Search 🕒 UTC ✅ Done (0.053 s) 📄 5 records 🔗

Warehouse_block	Count1	Count0	Percentage1	Percentage0
> F	2,194	1,472	59.85	40.15
> D	1,096	738	59.76	40.24
> B	1,104	729	60.23	39.77
> C	1,094	739	59.68	40.32
> A	1,075	758	58.65	41.35



Conclusion: The Warehouse Block with the best on time delivery is B at 60.23%. In descending order the rest are F at 59.85%, D at 59.76%, C at 59.68%, and A at 58.65%.

Question 4: Which is each warehouse block's percentage of on time delivery, which warehouse block is the best? (Power BI)



```
1 Min_OnTime = MIN('Train'[Reached.on.Time_Y.N])
```

```
OnTimePercentage =  
DIVIDE(  
    CALCULATE(SUM('Train'[Reached.on.Time_Y.N])),  
    CALCULATE(COUNTROWS('Train'))  
) * 100
```

Visualization 1 (Top Left)

Shows the Percentage of Packages arriving on time based on what Warehouse Block It came from.

Visualization 2 (Top Right)

Compares Warehouse Block F to the rest of the warehouse blocks.

Visualization 3 (Bottom Left)

Compares product_importance high to the arrival on time percentage at all warehouse blocks in the system.

Visualization 4 (Bottom Right)

Compares 1= yes which is if the package arrived on time to the rest of the warehouse blocks.

Question 5: What is the average customer rating for each warehouse block? (Azure)

```
1 dataset
2 | summarize Average_Customer_Rating = avg(Customer_rating) by Warehouse_block
3 | project Warehouse_block, Average_Customer_Rating
4
```

Table 1 + Add visual Stats Search UTC Done (0.129 s) 123 5 records

Warehouse_block	Average_Customer_Rating
F	2.9967266775777412
D	3.0179934569247546
B	2.9803600654664484
C	2.993998908892526
A	2.9574468085106385

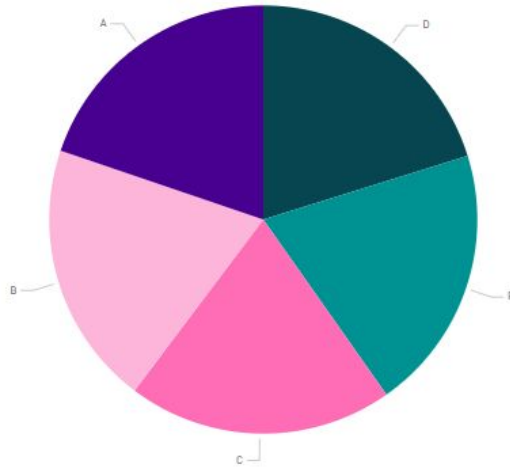


Conclusion: The average customer rating for each warehouse block is very similar. The highest rating is D with an average of 3.017, then F with an average of 2.996, then C with an average of 2.993, then B with an average of 2.980, then A with an average of 2.957.

Question 5: What is the average customer rating for each warehouse block? (Power BI)

```
AvgCustomerRating = AVERAGE('Train'[Customer_rating])
```

AvgCustomerRating by Warehouse_block



Conclusion:
The highest average customer rating is D followed by F, followed by C, followed by B, followed by A

Question 6: What ethical considerations should E-Commerce businesses consider when using customer data?

Data is important for showing how customers shop, their experience shopping, and the potential pain points in the shopping process. These factors make data one of the major backbones of E-Commerce business. Because of the necessity of data in E-Commerce business using that data in an ethical way becomes far more important. There are several guidelines an E-Commerce business should incorporate to ensure the ethical use of data. One guideline is transparency. This guideline in a data policy shows what data is being collected and the purpose of why it is being collected. Another guideline is honesty. This guideline in a data policy ensures the company is honest with what data they collect and the purpose of collection. Another guideline is relevancy. This guideline in a data policy ensures only relevant data is collected to provide less risk for the customer. The final guideline is security. This is the most important factor in data ethics. Unsafe security of data can lead to disruptions of a customer's life. When considering ethics in E-Commerce business security should be the first priority.

Conclusion

This project provided our group with new information about Microsoft Azure, Power BI, and Chat GPT. We learned how to use the natural language model of Chat GPT to generate multiple different kinds of code for use in Azure or Power BI to produce the desired results in our dataset. We learned the major strengths and weaknesses of each program and in what circumstances it would be appropriate to use each program. Finally, we further educated ourselves on ethical data practices and the importance of it.

References

<https://www.kaggle.com/datasets/prachi13/customer-analytics>

<https://www.paymentsjournal.com/ethical-guidelines-for-the-use-of-e-commerce-data/>

<https://chat.openai.com/>

<https://azure.microsoft.com/en-us>

<https://dataexplorer.azure.com/>

<https://powerbi.microsoft.com/en-us/>