Microsoft Azure Project

Alanna Hartzell and Rachel Vlassis

Our Azure Training

Alanna:

- Fundamentals of Azure AI Services
- Fundamentals of Machine Learning
- Explore fundamentals of of data visualization
- Use Azure Resource Manager
- Introduction to Azure OpenAI Service
- Use Automated Machine Learning in Azure ML
- Explore fundamentals of real-time analytics

Rachel:

- Microsoft Azure Fundamentals: Explore core data concepts
- Microsoft Azure Data Fundamentals: Explore relational data in Azure
- Microsoft Azure Data Fundamentals: Explore non-relational data in Azure
- Microsoft Azure Data Fundamentals: Explore data analytics in Azure
- Get Started with Power BI

Our Dataset

Our dataset consists of 8,763 records, each corresponding to a patient from across the globe. 24 different attributes were collected on each patient relating to their heart health in order to determine their current risk for a heart attack. We retrieved our dataset from Kaggle.com, it can be found here.

A preview of our dataset can be found below. It includes a unique identifier for each patient and demographics such as age, sex, income, country, continent, and hemisphere. It also includes 17 different patient-specific details ranging from blood pressure to sleep hours per day. These details were used to inform the last column, heart attack risk. Based on the information provided by the patients, it was determined if they are at risk for a heart attack, indicated by a 1, or they are not, indicated by a 0.

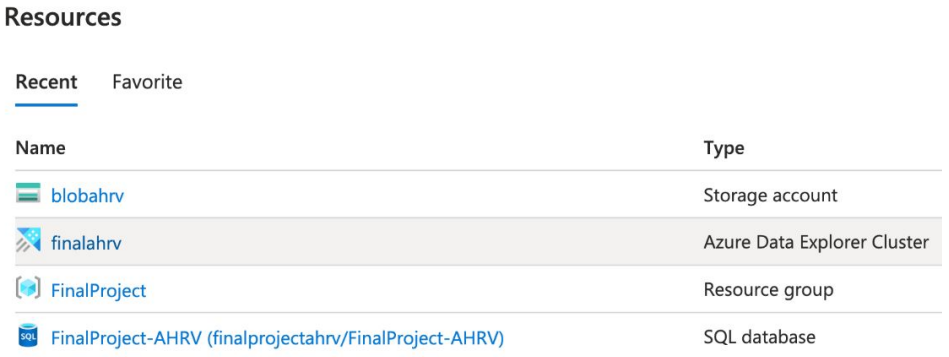
We chose this dataset as heart disease is the leading cause of death globally and affects people in our own lives. As data analysts, it is our responsibility to use our skills to transform raw data such as this into information that can be used to educate individuals on decreasing their risk for developing heart disease.

Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Diet	Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day	Income	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Country	Continent	Hemisphere	Heart Attack Risk
HQM9364	75	Female	136	141/85	101	No	No	0	1	Light	14.744881	Unhealthy	0	1	4	10.922177	94152	30.589796	374	3	4	Nigeria	Africa	Northern Hemisphere	1
ESJ9954	62	Male	262	137/82	89	Yes	No	0	1	Light	16.228489	Unhealthy	0	1	7	1.208610	159792	31.584511	678	0	8	Australia	Australia	Southern Hemisphere	1
ONA1218	72	Female	126	138/93	86	Yes	Yes	1	0	None	6.818887	Average	0	1	4	9.514556	254952	34.711478	736	1	5	Argentina	South America	Southern Hemisphere	1
UBE5339	18	Female	300	132/94	109	Yes	No	1	1	Light	18.297860	Average	1	1	6	9.015221	25229	29.022289	152	2	5	Spain	Europe	Southern Hemisphere	1
LUQ7367	67	Female	223	91/89	84	Yes	Yes	1	0	Moderate	10.980701	Average	0	1	4	10.020410	229179	35.966244	744	5	8	Japan	Asia	Northern Hemisphere	1

Using Azure Portal

Steps Taken to Successfully Upload Data:

1. Created SQL database in Azure
    a. This was not useful. We could not upload data into it.
2. Created Blob Storage Account in Azure
3. Uploaded dataset into Blob
4. Created Azure Data Explorer Cluster
5. Linked Blob with dataset to Azure Explorer



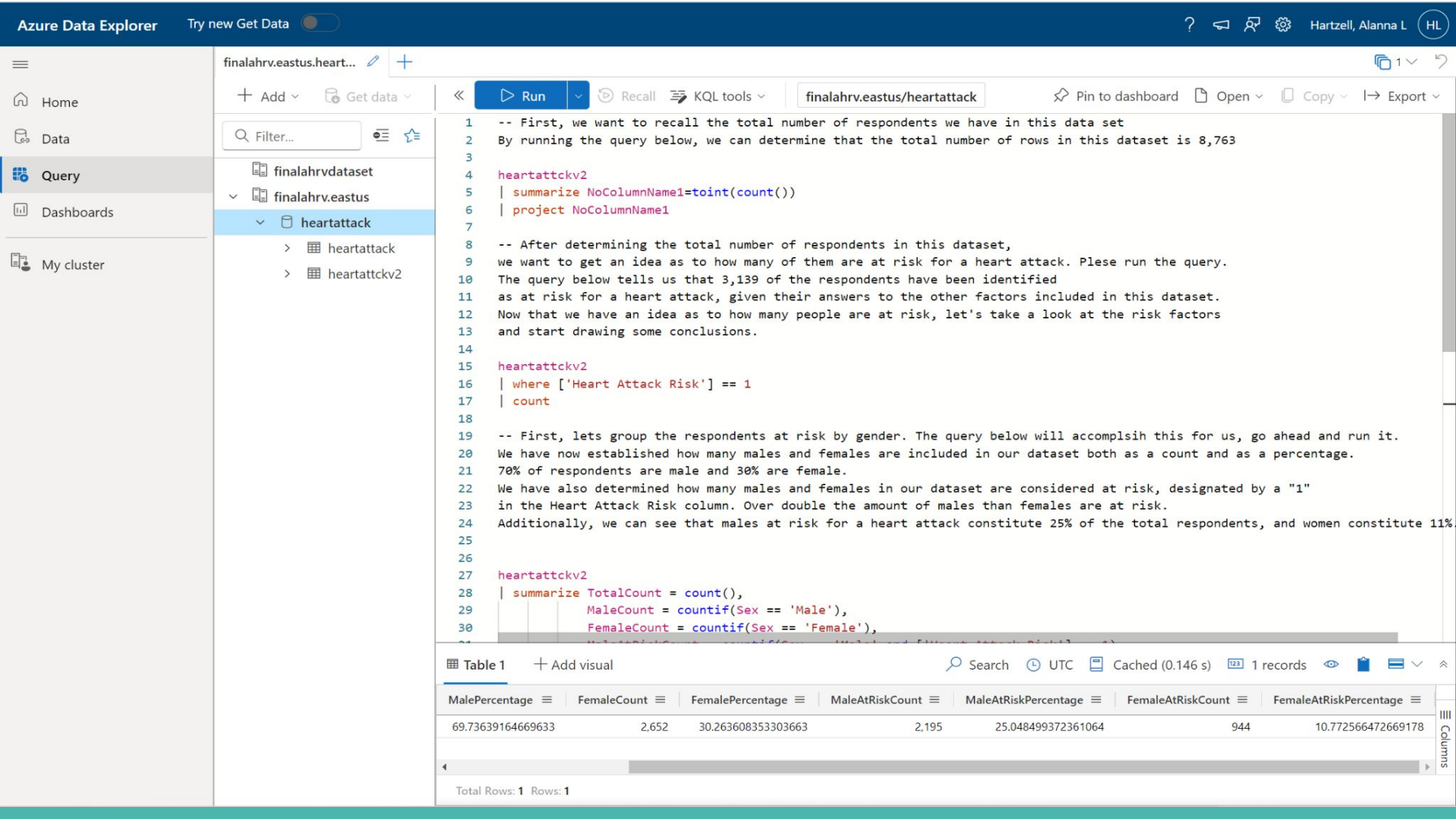
KQL vs SQL

Azure Data Explorer utilizes **KQL**, or **Kutso Query Language**. KQL is designed for querying large volumes of structured and semi-structured data. Prior to this course we were only familiar with **SQL**, which is used to query large volumes of structured data stored within relational databases. In order to accomplish the task of using Azure Data Explorer and querying in KQL, we first wrote our code in SQL since it is the language we know, and then used ChatGPT to help us turn our SQL queries into useable KQL queries.

Our Queries

In developing our data story, we wanted to start with the basics to understand the patients represented in our dataset before moving onto more complex questions. Over the next three slides, you will see a series of queries written in **KQL** within Azure Data Explorer as well as explanations prior to each query of what it will be performing. We sought to answer the following questions:

1. How many total records, or patients, are in our dataset?
2. How many of the patients are at risk for a heart attack?
3. How many males and females are in our dataset? How do they compare in terms of their risk for a heart attack?


Azure Data Explorer
Try new Get Data
Hartzell, Alanna L
HL
finalahrv.eastus.heart...
Home
Data
Query
Dashboards
My cluster
Add
Get data
Run
Recall
KQL tools
finalahrv.eastus/heartattack
Pin to dashboard
Open
Copy
Export
Filter...
finalahrvdataset
finalahrv.eastus
heartattack
heartattack
heartattckv2
-- First, we want to recall the total number of respondents we have in this data set
By running the query below, we can determine that the total number of rows in this dataset is 8,763
heartattckv2
| summarize NoColumnName1=toint(count())
| project NoColumnName1
-- After determining the total number of respondents in this dataset,
we want to get an idea as to how many of them are at risk for a heart attack. Plese run the query.
The query below tells us that 3,139 of the respondents have been identified
as at risk for a heart attack, given their answers to the other factors included in this dataset.
Now that we have an idea as to how many people are at risk, let's take a look at the risk factors
and start drawing some conclusions.
heartattckv2
| where ['Heart Attack Risk'] == 1
| count
-- First, lets group the respondents at risk by gender. The query below will accomplsih this for us, go ahead and run it.
We have now established how many males and females are included in our dataset both as a count and as a percentage.
70% of respondents are male and 30% are female.
We have also determined how many males and females in our dataset are considered at risk, designated by a "1"
in the Heart Attack Risk column. Over double the amount of males than females are at risk.
Additionally, we can see that males at risk for a heart attack constitute 25% of the total respondents, and women constitute 11%.
heartattckv2
| summarize TotalCount = count(),
MaleCount = countif(Sex == 'Male'),
FemaleCount = countif(Sex == 'Female'),
Table 1
Add visual
Search
UTC
Cached (0.146 s)
1 records
MalePercentage
FemaleCount
FemalePercentage
MaleAtRiskCount
MaleAtRiskPercentage
FemaleAtRiskCount
FemaleAtRiskPercentage
69.73639164669633
2,652
30.263608353303663
2,195
25.048499372361064
944
10.772566472669178
Columns
Total Rows: 1 Rows: 1

Azure Data Explorer Try new Get Data Hartzell, Alanna L HL

Home
Data
Query
Dashboards
My cluster

finalahrv.eastus.heart...

Add Get data Run Recall KQL tools finalahrv.eastus/heartattack Pin to dashboard Open Copy Export

Filter...

- finalahrvdataset
- finalahrv.eastus
  - heartattack
    - heartattack
    - heartattckv2

```
-- After determining the total number of respondents in this dataset,
we want to get an idea as to how many of them are at risk for a heart attack. Plese run the query.
The query below tells us that 3,139 of the respondents have been identified
as at risk for a heart attack, given their answers to the other factors included in this dataset.
Now that we have an idea as to how many people are at risk, let's take a look at the risk factors
and start drawing some conclusions.

heartattckv2
| where ['Heart Attack Risk'] == 1
| count

-- First, lets group the respondents at risk by gender. The query below will accomplsih this for us, go ahead and run it.
We have now established how many males and females are included in our dataset both as a count and as a percentage.
70% of respondents are male and 30% are female.
We have also determined how many males and females in our dataset are considered at risk, designated by a "1"
in the Heart Attack Risk column. Over double the amount of males than females are at risk.
Additionally, we can see that males at risk for a heart attack constitute 25% of the total respondents, and women constitute 11%.


heartattckv2
| summarize TotalCount = count(),
            MaleCount = countif(Sex == 'Male'),
            FemaleCount = countif(Sex == 'Female'),
            MaleAtRiskCount = countif(Sex == 'Male' and ['Heart Attack Risk'] == 1),
            FemaleAtRiskCount = countif(Sex == 'Female' and ['Heart Attack Risk'] == 1)
| extend MalePercentage = toreal(MaleCount) * 100.0 / TotalCount,
         FemalePercentage = toreal(FemaleCount) * 100.0 / TotalCount,
         MaleAtRiskPercentage = toreal(MaleAtRiskCount) * 100.0 / TotalCount,
         FemaleAtRiskPercentage = toreal(FemaleAtRiskCount) * 100.0 / TotalCount
```

Table 1 Add visual Search UTC Cached (0.146 s) 1 records

TotalCount	MaleCount	MalePercentage	FemaleCount	FemalePercentage	MaleAtRiskCount	MaleAtRiskPercentage	FemaleAtRiskCo
8,763	6,111	69.73639164669633	2,652	30.263608353303663	2,195	25.048499372361064	

Columns

Total Rows: 1 Rows: 1

Azure Data Explorer Try new Get Data Hartzell, Alanna L HL

Home
Data
Query
Dashboards
My cluster

finalahrv.eastus.heart...

Add Get data Run Recall KQL tools finalahrv.eastus/heartattack Pin to dashboard Open Copy Export

Filter...

- finalahrvdataset
- finalahrv.eastus
  - heartattack
    - heartattack
    - heartattckv2

```
-- After determining the total number of respondents in this dataset,
we want to get an idea as to how many of them are at risk for a heart attack. Plese run the query.
The query below tells us that 3,139 of the respondents have been identified
as at risk for a heart attack, given their answers to the other factors included in this dataset.
Now that we have an idea as to how many people are at risk, let's take a look at the risk factors
and start drawing some conclusions.

heartattckv2
| where ['Heart Attack Risk'] == 1
| count

-- First, lets group the respondents at risk by gender. The query below will accomplsih this for us, go ahead and run it.
We have now established how many males and females are included in our dataset both as a count and as a percentage.
70% of respondents are male and 30% are female.
We have also determined how many males and females in our dataset are considered at risk, designated by a "1"
in the Heart Attack Risk column. Over double the amount of males than females are at risk.
Additionally, we can see that males at risk for a heart attack constitute 25% of the total respondents, and women constitute 11%.


heartattckv2
| summarize TotalCount = count(),
            MaleCount = countif(Sex == 'Male'),
            FemaleCount = countif(Sex == 'Female'),
            MaleAtRiskCount = countif(Sex == 'Male' and ['Heart Attack Risk'] == 1),
            FemaleAtRiskCount = countif(Sex == 'Female' and ['Heart Attack Risk'] == 1)
| extend MalePercentage = toreal(MaleCount) * 100.0 / TotalCount,
         FemalePercentage = toreal(FemaleCount) * 100.0 / TotalCount,
         MaleAtRiskPercentage = toreal(MaleAtRiskCount) * 100.0 / TotalCount,
         FemaleAtRiskPercentage = toreal(FemaleAtRiskCount) * 100.0 / TotalCount
```

Table 1 Add visual Search UTC Cached (0.146 s) 1 records

MalePercentage	FemaleCount	FemalePercentage	MaleAtRiskCount	MaleAtRiskPercentage	FemaleAtRiskCount	FemaleAtRiskPercentage
69.73639164669633	2,652	30.263608353303663	2,195	25.048499372361064	944	10.772566472669178

Columns

Total Rows: 1 Rows: 1

Understanding our Query Results

1. How many total records, or patients, are in our dataset?

   8,763

2. How many of the patients are at risk for a heart attack?

   3,139

3. How many males and females are in our dataset? How do they compare in terms of their risk for a heart attack?

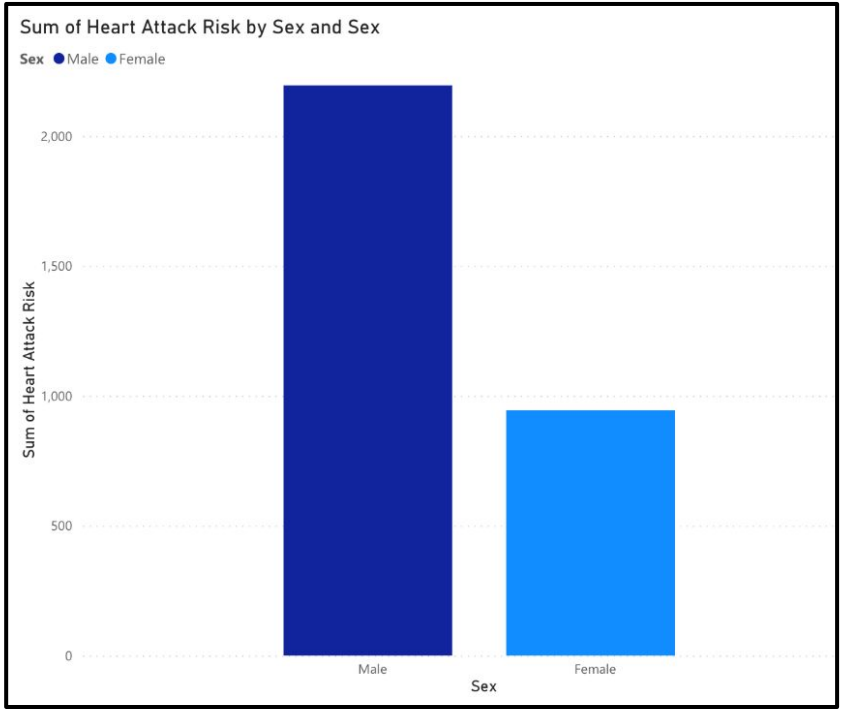
   There are 6,111 males and 2,652 females in our dataset. Males constitute roughly 70% of the records, and females 30%. Of the 3,139 patients at risk, 2,195 of them are male and 944 are female. It is important to note that this number is over double, however so is male representation at 70% vs female at 30%. Additionally, we determined that males at risk for a heart attack constitute 25% of the total records, and females constitute 10%.

Our Visualizations in Power BI

Sex	Sum of Heart Attack Risk
Male	2195
Female	944

Because our Azure accounts have been disabled, We were unable to export our completed codes to PowerBI that we had planned. Therefore, we attempted to manually create visualizations related to our queries. They are not as detailed as we would have liked them to be, but PowerBI does not offer many tools to filter out data in the way that coding does.

PowerBI allows you to input a dataset and it will create a report, in other words an overview of the data. This was helpful for us to get an understanding of the data and numbers were working with since we were unable to continue coding on our Azure accounts. The following slide shows this report.

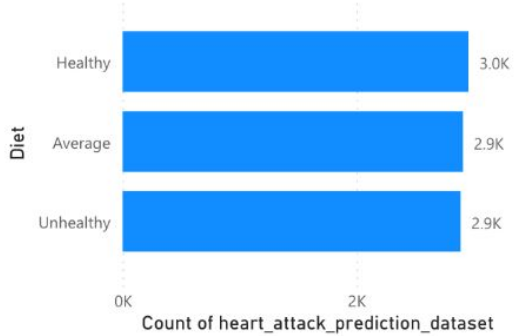


Quick summary

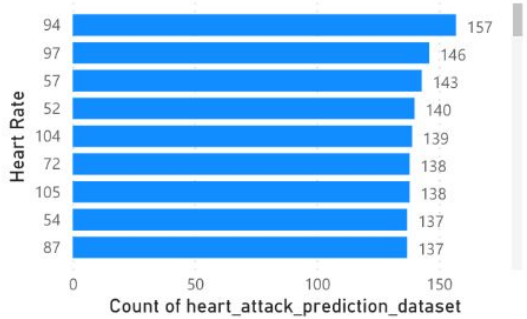
heart_attack_prediction_dataset

8763
Count of heart_attack_pr...

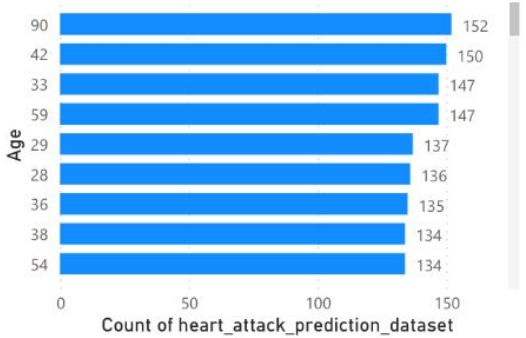
Count of heart_attack_prediction_dataset by Diet



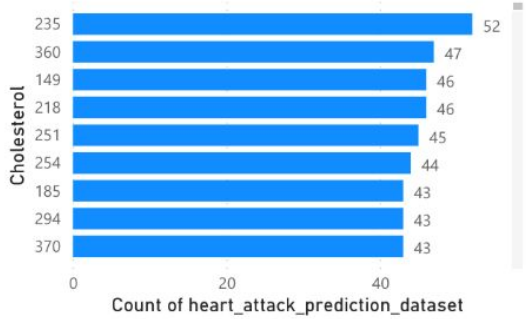
Count of heart_attack_prediction_dataset by Heart Rate



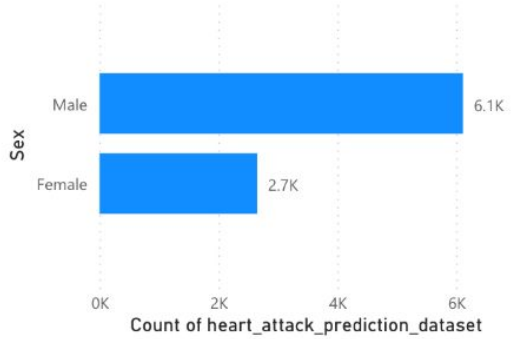
Count of heart_attack_prediction_dataset by Age



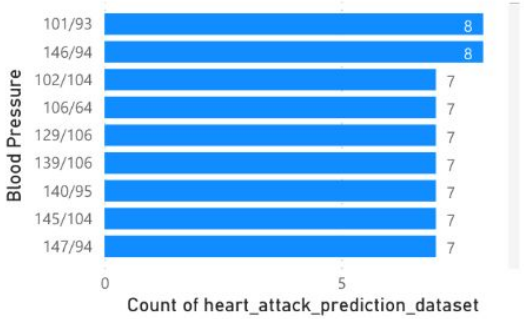
Count of heart_attack_prediction_dataset by Cholesterol



Count of heart_attack_prediction_dataset by Sex



Count of heart_attack_prediction_dataset by Blood Pressure



Our Azure Experience

After reviewing the queries that we ran, we really only scratched the surface of this dataset. We aimed to develop more questions to gather more insight into this dataset, but our ability to code and run queries in Azure Data Explorer was cut short when both of our licenses were unfortunately disabled. The disabling over our accounts prevents us from executing any further queries in Azure Data Explorer.

Overall, we were still able to successfully train in Azure, navigate and utilize services in Azure Portal, upload our dataset into Azure Data Explorer, and run KQL in order to obtain information about our dataset.

In the proceeding slides we develop what the rest of our data story would look like if we had the ability to continue coding and utilizing Azure Data Explorer.

Continuing Our Data Story

Here are some more questions we would have liked to answer using our data set, as well as their KQL code:

What is the patient age range? What is the frequency of each age?

These questions helps us to continue understanding patient demographics so we have an understanding of the gender breakdown and age breakdown of the patients.

```
heartattckv2
| summarize MinAge = min(Age), MaxAge = max(Age),
        AgeFrequency = datatable(Age: int, Frequency: int)
                [Age]
                | summarize Frequency = count() by Age
```

Data Story continued

What are the average cholesterol, blood pressure, and heart rate the patients determined to be at risk for a heart attack?

Now that we have an idea for some patient demographics, it's time to get to get to the heart of the dataset. This question allows us to begin exploring the dynamic between risk factors and if the patient is or is not at risk for a heart attack. Cholesterol, blood pressure, and heart rate are commonly used as indicators for heart disease so it is important that patients understand what healthy vs unhealthy levels look like. Determining the average from the dataset helps us to relay what levels could be cause for concern.

```
heartattckv2

| where [Heart Attack Risk] == 1

| summarize AvgCholesterol = avg(Cholesterol),

        AvgBloodPressure = avg([Blood Pressure]),

        AvgHeartRate = avg([Heart Rate])
```

Data Story continued

What are the average exercise hours per week, stress level, and sleep hours for patients determined to be at risk for a heart attack? What about the patients that are not at risk?

After looking at cholesterol, blood pressure and heart rate which are physical findings related to heart health, we also want to get insight into the patient reported findings such as exercise hours per week, stress level, and sleep hours. It is important to understand what these variables look like for both those at risk and those not at risk so that patients can be educated on how to create heart healthy habits and what they should avoid.

```
heartattckv2

| summarize AvgExerciseHoursAtRisk = avg(iff([Heart Attack Risk] == 1, [Exercise Hours Per Week], null)),

        AvgStressLevelAtRisk = avg(iff([Heart Attack Risk] == 1, [Stress Level], null)),

        AvgSleepHoursAtRisk = avg(iff([Heart Attack Risk] == 1, [Sleep Hours Per Day], null)),

        AvgExerciseHoursNotAtRisk = avg(iff([Heart Attack Risk] == 0, [Exercise Hours Per Week], null)),

        AvgStressLevelNotAtRisk = avg(iff([Heart Attack Risk] == 0, [Stress Level], null)),

        AvgSleepHoursNotAtRisk = avg(iff([Heart Attack Risk] == 0, [Sleep Hours Per Day], null))
```

Data Story continued

What country has the highest number of patients determined to be at risk for a heart attack? What country has the lowest?

In wrapping up our questions, we thought it would be important to look at heart attack risk by country to see which reporting country has the highest number of patients determined to be at risk, and which has the lowest. Habits and lifestyles are certainly linked to culture, and where we live and the food we eat has a significant impact on our overall health. It would be interesting to find a dataset that has information such as economic indicators or perhaps healthcare data to gain a better understanding as to how the two countries compare.

```
heartattckv2

| where [Heart Attack Risk] == 1

| summarize TotalAtRisk = count() by Country

| top 1 by TotalAtRisk desc
```

```
heartattckv2

| where [Heart Attack Risk] == 0

| summarize TotalAtRisk = count() by Country

| top 1 by TotalAtRisk desc
```

Thank You For Learning About Our Azure Experience