Group Members:

1. Sadaf Davre
2. Sharvil Shastri

INST754: Project Checkpoint

We have taken a strategic pivot in our project to better address the challenge of enhancing diabetes prevention and management in the American population. Just as a reminder, this is our original business problem and question:

Business Challenge: Enhance Diabetes Prevention and Management

Our Business Challenge is to analyze and determine which medical conditions play a significant role and are found in patients prone to diabetes. We will be able to suggest to those patients whether they should get tested for diabetes or not, based on pre-existing medical conditions. We will leverage a dataset to develop reports and analytics that can help identify individuals at risk of developing diabetes and provide insights for better patient care.

The following outlines the recent modifications and progress:

Tasks Completed:

Dataset Transition and Download: We transitioned to the "Behavioral Risk Factor Surveillance System (BRFSS)" dataset, which surveys over 400,000 people from 2011-2015, provided by the ***CDC***. This dataset is much more comprehensive and was selected for its broader range of health-related behaviors, chronic health conditions, and preventive service use.

Source Data Set: "Behavioral Risk Factor Surveillance System : Public health surveys of 400k people from 2011-2015"

https://www.cdc.gov/brfss/annual_data/annual_data.htm

ETL Process Initiation: Began the Extract, Transform, and Load (ETL) processes using Power Query within Power BI to accommodate the dataset's size and complexity. This step was crucial to prepare the data for insightful analysis.

Data Cleaning and preparation: Have performed initial data cleaning, which involved purging null values, expunging irrelevant data, standardizing column names for clarity, and translating numerical range values into a more interpretable format. The 2015 BRFSS data, consisting of 330 columns and 441,456 rows, became our focal point from all the years provided in the original set. We embarked on a meticulous data cleaning process for the same using Power Query and Power BI, optimizing the dataset for our analytical needs.

This included:

- Removing null values and irrelevant data
- Renaming columns for clarity.
- Converting numerical ranges to interpretable values.

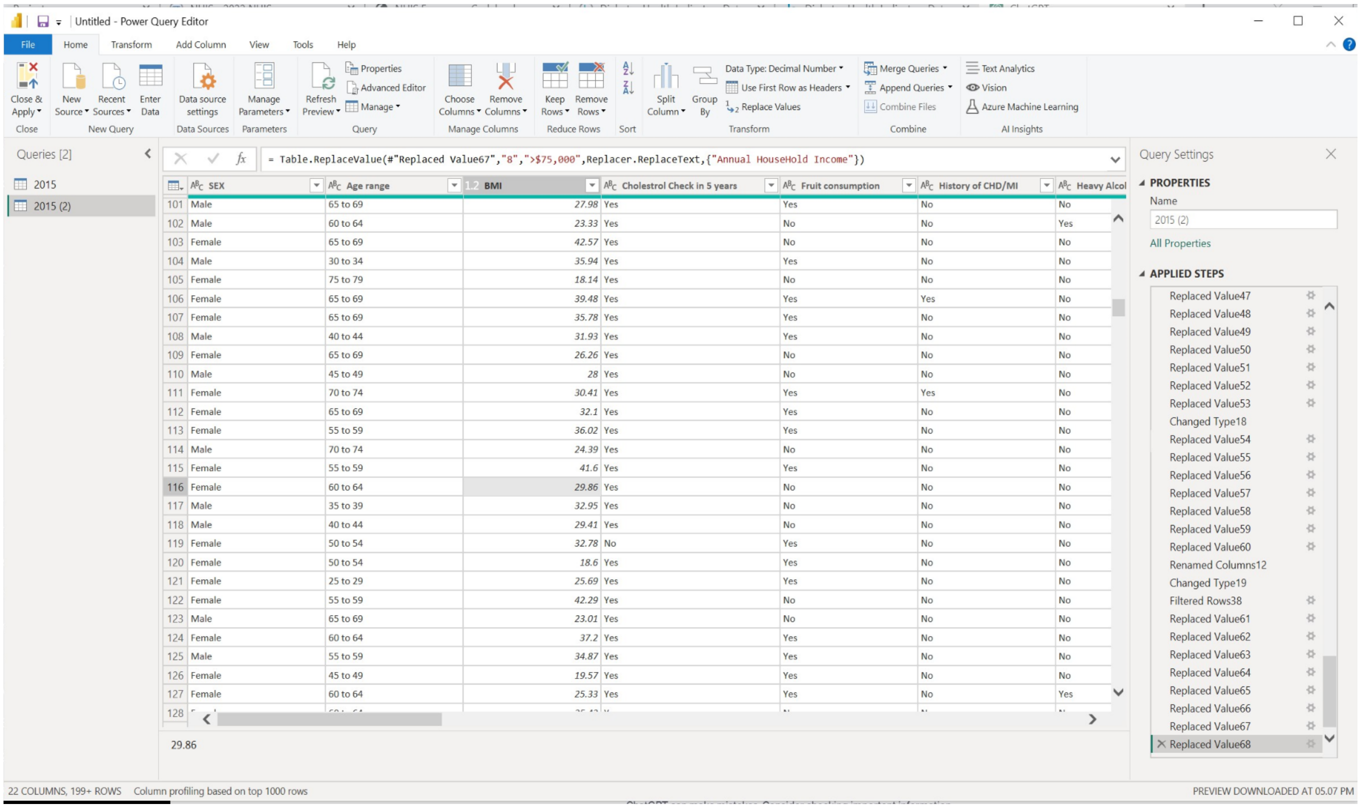
Report Requirements Definition: Outlined the specifications for the analytical reports and dashboards to be generated in Power BI, intending to highlight significant health factors influencing diabetes risk.

Feature Selection Based on Risk Factors: Informed by research on diabetes and chronic illnesses, we narrowed down our analysis to key risk factors such as high blood pressure, cholesterol levels, smoking habits, and obesity. The BRFSS 2015 Codebook was utilized to decode the dataset's variables.The codebook can be found here :

BRFSS 2015 Codebook: https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Literature Consultation for Feature Insight: Consulted academic literature like that of Zidian Xie et al., to guide our feature selection, to inform and validate the selection of features that are critical to understanding diabetes risk factors. Relevant Research Paper using BRFSS for Diabetes ML: https://www.cdc.gov/pcd/issues/2019/19_0109.htm

Data Structuring and Preparation:Post-cleaning, we structured the data, now comprising 253,680 survey responses, and 22 columns, into a format suitable for working in PowerBI to identify the likelihood of diabetes among the American population. This is a snapshot of our work done in PowerBI:



```
= Table.ReplaceValue(#"Replaced Value67","8",">$75,000",Replacer.ReplaceText,{"Annual HouseHold Income"})
```

	SEX	Age range	BMI	Cholestrol Check in 5 years	Fruit consumption	History of CHD/MI	Heavy Alcol
101	Male	65 to 69	27.98	Yes	Yes	No	No
102	Male	60 to 64	23.33	Yes	No	No	Yes
103	Female	65 to 69	42.57	Yes	No	No	No
104	Male	30 to 34	35.94	Yes	Yes	No	No
105	Female	75 to 79	18.14	Yes	No	No	No
106	Female	65 to 69	39.48	Yes	Yes	Yes	No
107	Female	65 to 69	35.78	Yes	Yes	No	No
108	Male	40 to 44	31.93	Yes	Yes	No	No
109	Female	65 to 69	26.26	Yes	No	No	No
110	Male	45 to 49	28	Yes	No	No	No
111	Female	70 to 74	30.41	Yes	Yes	Yes	No
112	Female	65 to 69	32.1	Yes	Yes	No	No
113	Female	55 to 59	36.02	Yes	Yes	No	No
114	Male	70 to 74	24.39	Yes	No	No	No
115	Female	55 to 59	41.6	Yes	Yes	No	No
116	Female	60 to 64	29.86	Yes	No	No	No
117	Male	35 to 39	32.95	Yes	No	No	No
118	Male	40 to 44	29.41	Yes	No	No	No
119	Female	50 to 54	32.78	No	Yes	No	No
120	Female	50 to 54	18.6	Yes	Yes	No	No
121	Female	25 to 29	25.69	Yes	Yes	No	No
122	Female	55 to 59	42.29	Yes	No	No	No
123	Male	65 to 69	23.01	Yes	No	No	No
124	Female	60 to 64	37.2	Yes	Yes	No	No
125	Male	55 to 59	34.87	Yes	Yes	No	No
126	Female	45 to 49	19.57	Yes	Yes	No	No
127	Female	60 to 64	25.33	Yes	Yes	No	Yes

29.86

Replaced Value47
Replaced Value48
Replaced Value49
Replaced Value50
Replaced Value51
Replaced Value52
Replaced Value53
Changed Type18
Replaced Value54
Replaced Value55
Replaced Value56
Replaced Value57
Replaced Value58
Replaced Value59
Replaced Value60
Renamed Columns12
Changed Type19
Filtered Rows38
Replaced Value61
Replaced Value62
Replaced Value63
Replaced Value64
Replaced Value65
Replaced Value66
Replaced Value67
Replaced Value68

22 COLUMNS, 199+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 05.07 PM

Upcoming Tasks and Goals:

1. **Descriptive Analytics Implementation:** Plan to utilize descriptive analytics within Power BI to discern patterns and relationships within the risk factors for diabetes.
2. **Power BI Report Development:** Develop detailed reports and interactive dashboards in Power BI to illustrate our findings regarding the prevalence and risk factors associated with diabetes.
3. **Final Presentation and Documentation:** Prepare a final presentation of our analytical process and findings, alongside documentation that outlines our methodologies and results.

Challenges encountered and mitigation strategies:

1. **Adapting to a New Dataset:** After shifting from the initial dataset to the BRFSS dataset, we faced the challenge of familiarizing ourselves with a new data structure and a broader set of variables. To cope with this, we have been meticulously studying the

BRFSS 2015 Codebook and aligning our understanding with the dataset's framework to ensure accurate analysis.

2. **Data Integrity and Quality Assurance:** Ensuring the integrity of the data after the initial cleaning has been paramount. To maintain high data quality, we have implemented validation checks and used Power BI’s data modeling features to confirm that our dataset is accurate and reliable for analysis.
3. **Time Management:** Balancing the project's extensive scope within the allotted time frame has required careful planning. We are addressing this by concentrating on essential tasks that will contribute significantly to the analysis and report generation, ensuring efficient use of our resources and time.
4. **Visualization Limitations in Power BI:** Although Power BI is a powerful tool for data visualization, there is a possibility that it may not have certain specific visuals we require. If we encounter such limitations, we will explore the wide range of custom visuals available within the Power BI community or adapt our visualization strategy to softwares like Tableau.