D211: Advanced Data Acquisition Task 1

Justin Huynh

Student ID: 012229514

M.S. Data Analytics

A. Data Dashboards

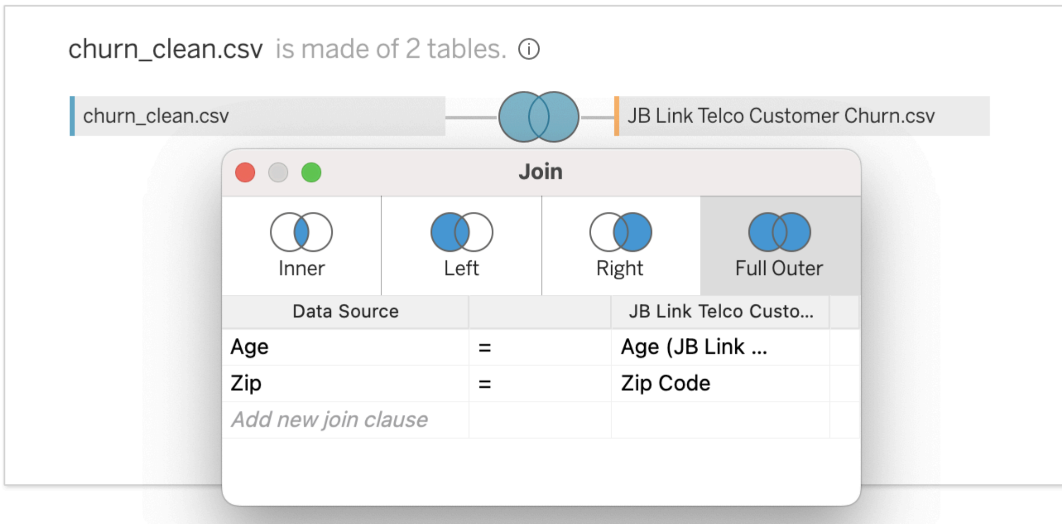
The dashboard will be submitted with the task.

A1. Data Sets and Dashboard Files

The two data sets being provided are the WGU churn data set (churn_clean.csv) and the JB Telco customer churn data set from Kaggle (JB Link Telco Customer Churn.csv). Both will be submitted in this task.

A2. Dashboard Installation

The dashboard will be submitted as a Tableau workbook labeled d211.twbx that a user can open directly for ease of use as long as Tableau is installed, which should be on Labs on Demand or on any local device that has Tableau on their desktop. Since the internal and external data sets were located on my desktop at the time of the Tableau dashboard creation, I believe the user may need to pull the data sets from their desktop as well after downloading both data sets submitted with the task. For simple installation, a user can download the internal and external data sets from the csv files submitted with this task, manually upload the internal and external data separately onto Tableau, and double click the data source to drag and drop the external with the internal files as a join to create a single data source file to pull the data from when creating the dashboard. This will allow the dashboard to populate and show the visualizations for the user to navigate through.



A3. Dashboard Navigation

Navigation of the dashboard is fairly simple. Once the dashboard has been opened up, they can view the story of the presentation, which has the 4 boxes at the top to navigate the different tabs. The intro tab provides a brief overview of what the presentation is about, the data sets used, and my name as the person who created the story.

The "Who Are The Customers?" tab is an interactive dashboard that shows the customer churn attributes by monthly charge, age, gender, and tenure. The data can be filtered in multiple ways. One being the age filter on the top left with the age slider for the WGU data set from 18-89 and another age slider for the JB Telco data set from 19-89 as there are the age ranges within their data sets. They will only update their respective data sets. The other filter is the average tenure by gender pie charts where they will update all the visualizations within the dashboard. Clicking on the gender of the pie chart will update the data on all visualizations and they can exit that view when clicking out of the chart so it goes back to default.

The “Conditions KPI” is another interactive dashboard that observes the metric as key performance indicators (KPI), which are churn rate by age and churn rate by gender, demonstrating what type of customers churn. In these bar charts, there will be filters for them as well with the age group filter for both the WGU and JB Telco data sets. When selecting the age group, the corresponding data set will update only on those visualizations. However, the gender bar charts have filters themselves and will update the entire dashboard when selecting the gender.

The last tab is the “Takeaways & Recommendations” where key takeaways will be given and recommendations will be provided based on the analysis of the dashboards. These, like the intro slide, are non-interactive so they will just be static texts.

A4. SQL Code

This is the SQL code inputted into PostgreSQL database to join and create the new single data set combined from the internal WGU churn and external data sets:

```
-- Join the external dataset with the location table on the zip columns with casting to
INTEGER
CREATE TABLE churn_jb_link_telco AS
WITH location_joined AS (
   SELECT
      j.zip_code,
      j.age AS churn_age, -- Rename this age column
      j.gender,
      j.monthly_charge,
```

```
        j.tenure_in_months,
        j.churn_value,
        l.city,
        l.state,
        l.county
    FROM
        jb_link_telco_customer_churn j
    JOIN
        location l
    ON
        j.zip_code::INTEGER = l.zip
)
-- Save the result as a new table named churn_jb_link_telco
SELECT
    c.customer_id,
    c.lat,
    c.lng,
    c.population,
    c.children,
    c.age AS customer_age, -- Rename this age column
    c.income,
    c.marital,
    c.churn,
```

```
    c.gender,
    c.tenure,
    c.monthly_charge,
    c.bandwidth_gp_year,
    c.outage_sec_week,
    c.email,
    c.contacts,
    c.yearly_equip_faiure,
    c.techie,
    c.port_modem,
    c.tablet,
    c.job_id,
    c.payment_id,
    c.contract_id,
    c.location_id,
    lj.churn_age, -- Use the renamed age column here
    lj.city,
    lj.state,
    lj.county
FROM
    customer c
JOIN
    location_joined lj
```

```
ON
   c.gender = lj.gender;
-- Copy relevant external data into PostgreSQL
COPY jb_link_telco_customer_churn(zip_code, age, gender, monthly_charge,
tenure_in_months, churn_value)
FROM '/Users/justinhuynh/Desktop/JB Link Telco Customer Churn.csv'
DELIMITER ','
CSV HEADER;
```

B: Panopto Presentation

A link to the Panotop video will be submitted with this task.

C: Written Report

The exploration phase involved understanding the structure, quality, and relevance of the data available in both internal and external datasets. The primary focus was on identifying key variables related to customer demographics, billing, and service usage that could influence churn behavior. Advanced SQL operations were central to transforming, aligning, and preparing the data for analysis. First, new tables were created in PostgreSQL to store and organize data from the external dataset with the SQL code in A4. The analysis phase focused on deriving insights related to customer churn to better understand what actionable actions can be taken. More information will be outlined below.

C1: Dashboard Alignment

The purpose of the dashboard is to provide stakeholders with a clear and actionable overview of customer churn patterns within the telecom industry. The dashboard aligns with stakeholders' needs by focusing on key metrics such as average monthly charges, customer tenure by months, and churn rates segmented by age and gender. These metrics are crucial for understanding customer behavior and identifying areas that can improve customer retention. The dashboard provides a high-level overview of churn trends, which is vital for making strategic decisions regarding pricing, service improvements, and customer engagement initiatives. It also allows teams to use insights from the dashboard to develop targeted retention strategies and identify customer segments that are more likely to churn.

C2: Business Intelligence Tool

Tableau was chosen as the business intelligence tool for this analysis due to its powerful data visualization capabilities, ease of use, and ability to handle complex data connections. It allows for the creation of dynamic, interactive visualizations that stakeholders can easily interpret and improve decision-making. Tableau also integrates seamlessly with various data sources, including PostgreSQL, which was used in this analysis. This integration ensures that data from multiple tables can be combined and visualized efficiently. The dashboard also includes interactive elements such as filters to enable stakeholders to explore the data in depth and focus on areas of interest. Lastly, Tableau is scalable and can handle large datasets, making it suitable for ongoing data analysis and reporting needs.

C3: Data Cleaning

The data cleaning and preparation process involved several steps to ensure that the data was ready for analysis. First, the data was standardized to ensure consistency, especially with columns like zip_code, where inconsistencies in data types were corrected through casting (e.g., from VARCHAR to INTEGER). Next, the COPY command was used to load data from external CSV files into the PostgreSQL database to ensure that all relevant data was available for analysis. Columns were then transformed to ensure compatibility between datasets. For example, the zip_code from the external dataset was cast to an integer to match the zip column in the internal location table. Duplicate columns (e.g., age in both the customer and jb_link_telco_customer_churn tables) were resolved by renaming or excluding one of the columns during joins. Lastly, data from multiple tables was joined using SQL JOIN operations based on common columns like zip. This process created a unified dataset that could be used to generate the dashboard in Tableau.

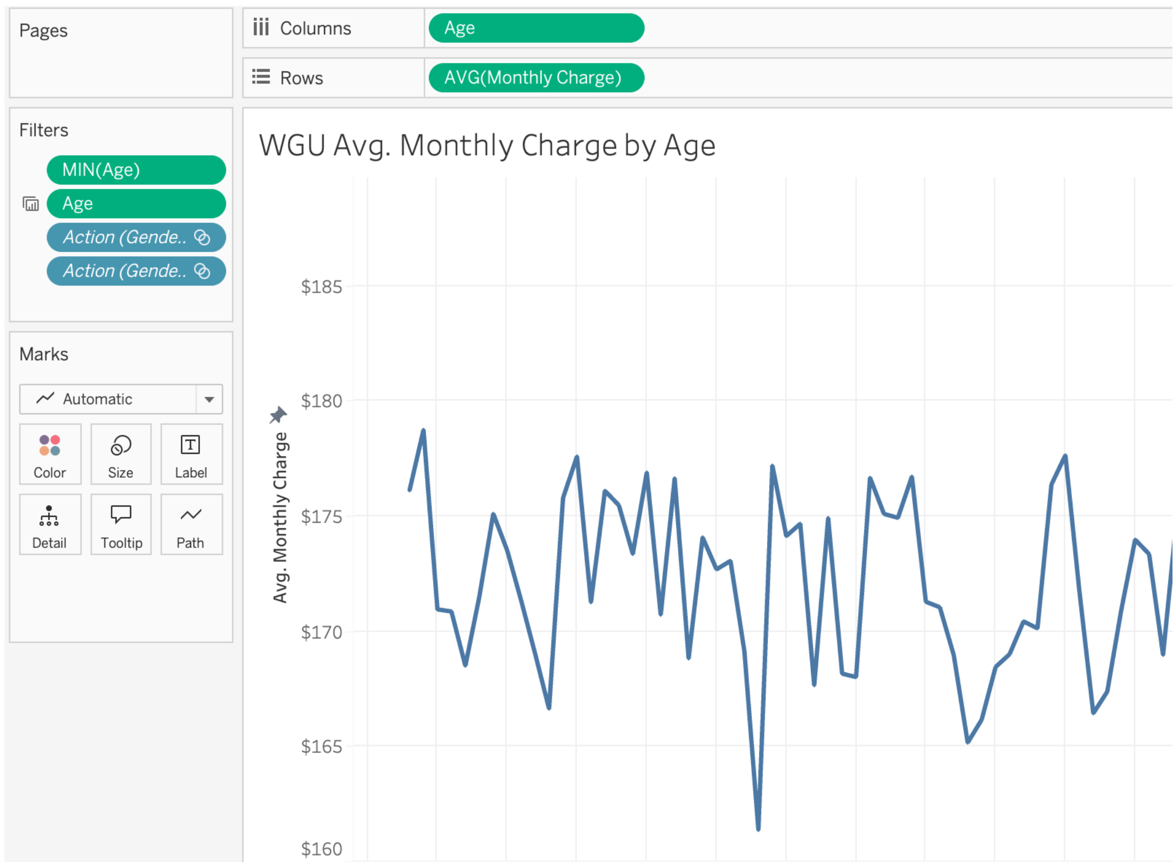
C4: Dashboard Creation

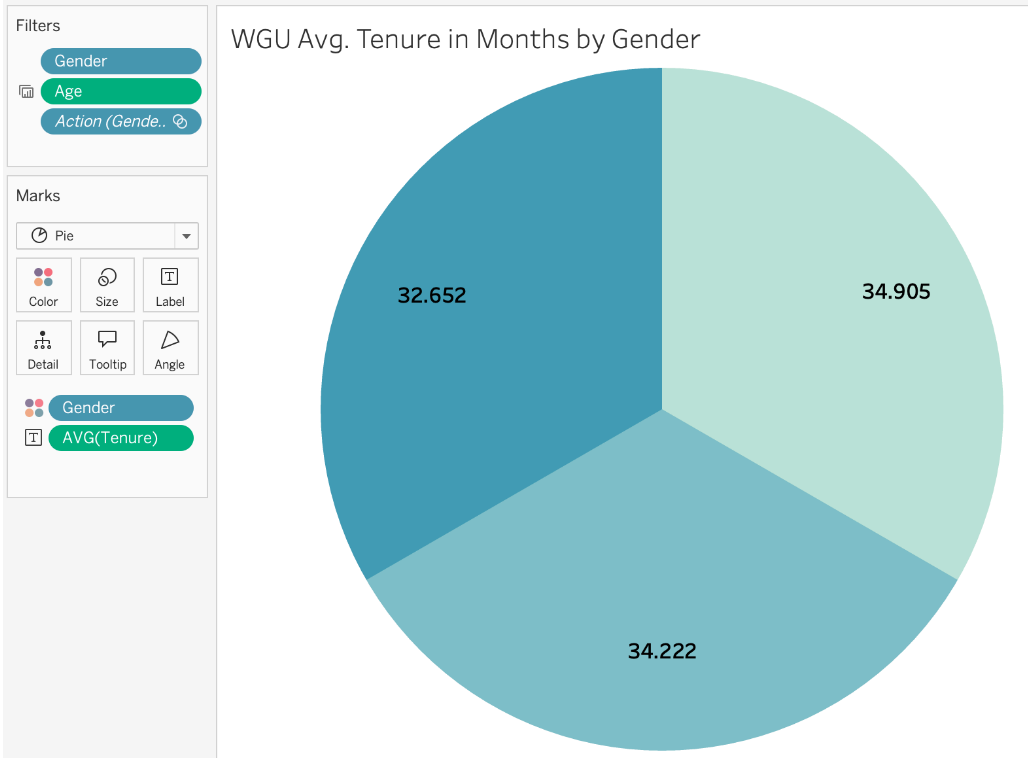
Since the dashboard is identical to D210, a copy was created for this project labeled D211.twb with the exception of the WGU and external datasets being joined into a single dataset and used as the primary data source vs. the dashboard pointing to 2 separate data sources. The creation of the dashboard involved several steps:

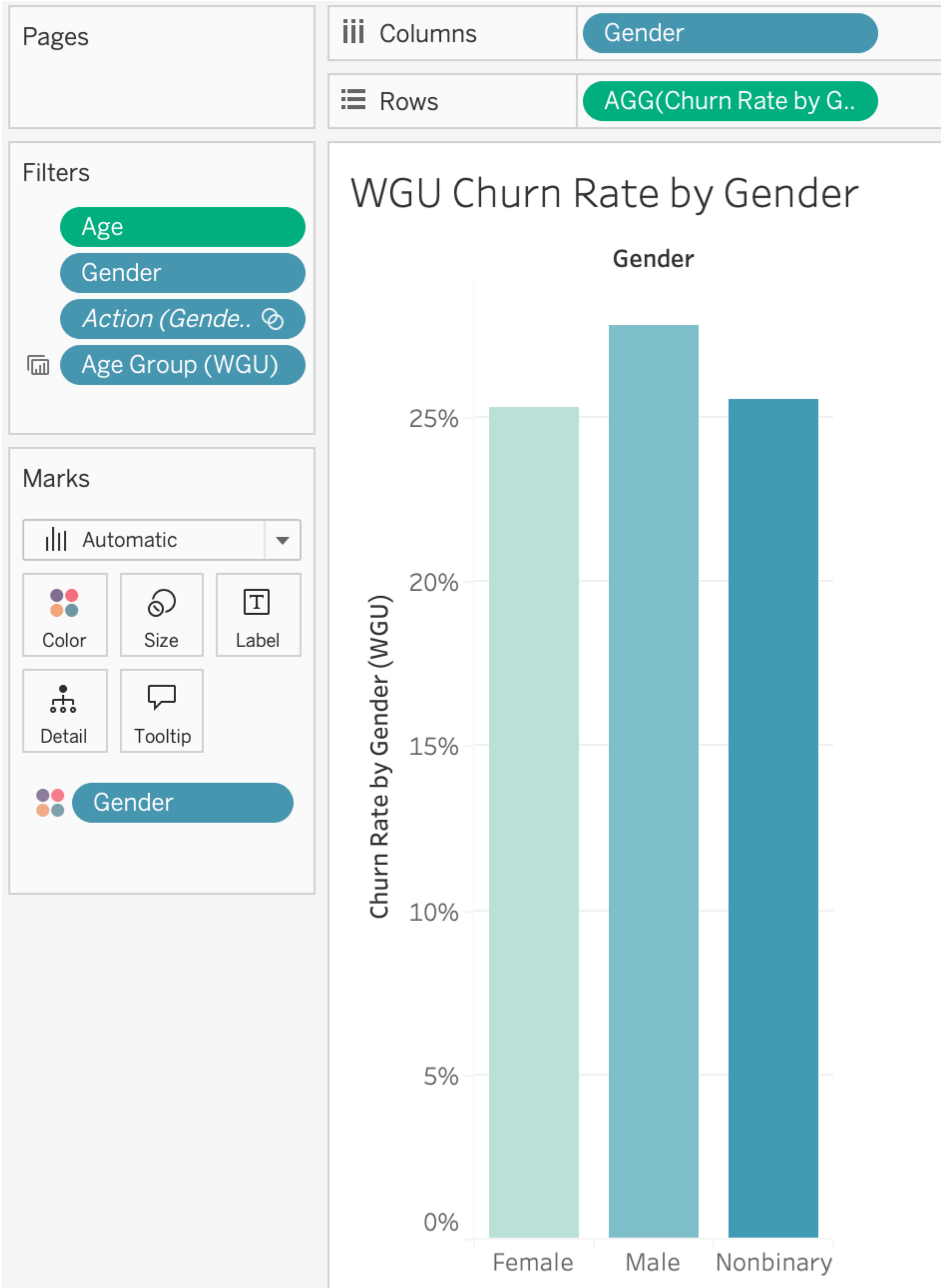
- Visualization Creation:
    - Various visualizations were created to represent key metrics such as average monthly charges, tenure by age and gender, and churn rates. Each visualization was designed to highlight important trends and provide actionable insights.
    - Example visualizations included line charts showing average monthly charges by age, pie charts depicting tenure by gender, and bar charts

illustrating churn rates by age group. The columns were dragged and dropped accordingly to create the visualizations.

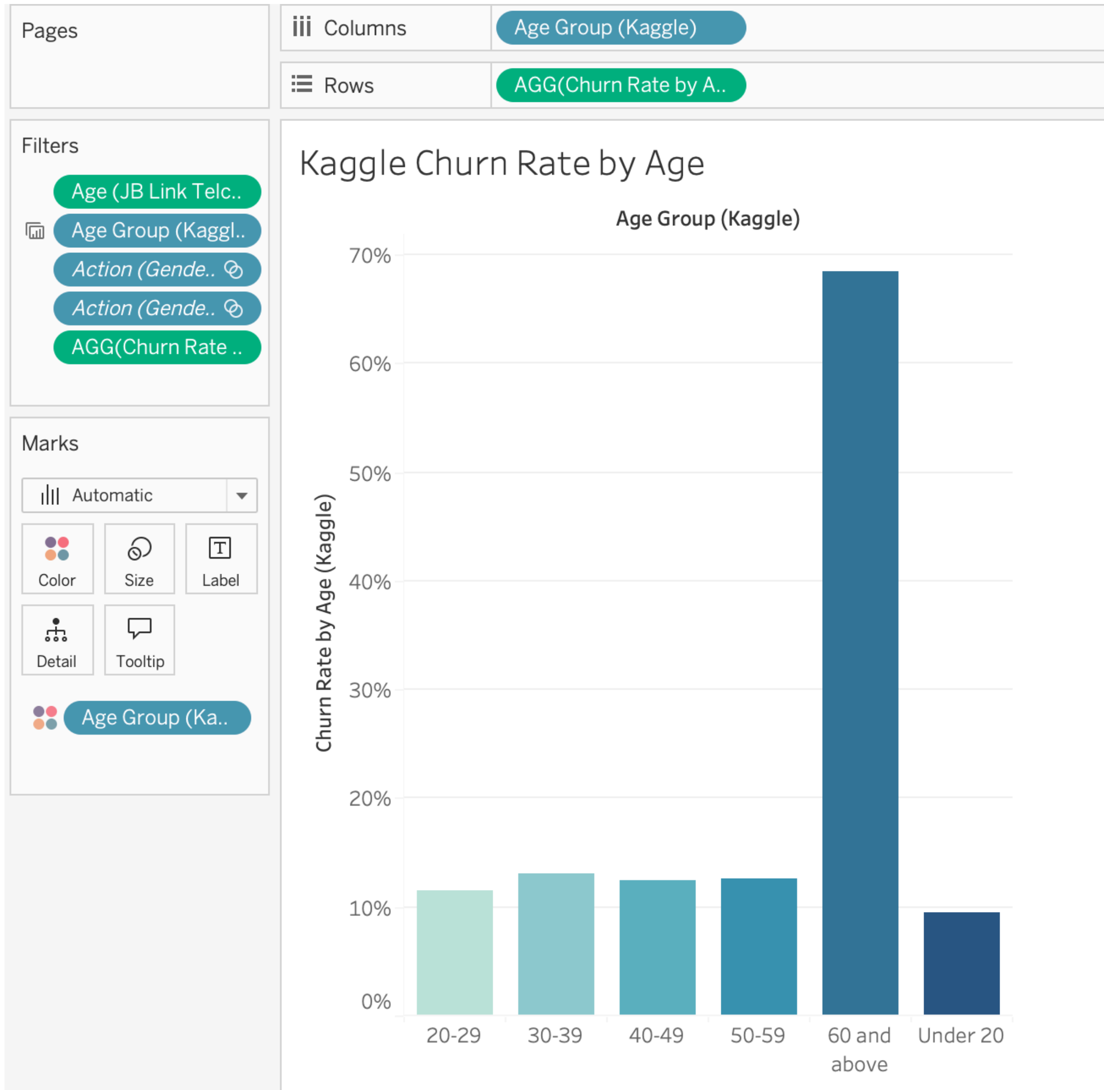
- Calculated fields were created for the conditional KPI's to create churn rate by age and gender.

- 

- 
Filters
Gender
Age
Action (Gende..
Marks
Pie
Color
Size
Label
Detail
Tooltip
Angle
Gender
AVG(Tenure)
WGU Avg. Tenure in Months by Gender
32.652
34.905
34.222


Pages
Columns
Gender
Rows
AGG(Churn Rate by G..
Filters
Age
Gender
Action (Gende..
Age Group (WGU)
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Gender
WGU Churn Rate by Gender
Gender
Churn Rate by Gender (WGU)
25%
20%
15%
10%
5%
0%
Female
Male
Nonbinary

○

○ 

- For the dashboard assembly, the individual visualizations were assembled into a comprehensive dashboard and filters and interactive elements were added to allow stakeholders to explore the data dynamically.
- A Tableau Story was created to guide stakeholders through the analysis, starting with an overview of the customer base, followed by condition KPIs, and concluding with key takeaways and recommendations.

C5: Data Analysis Results

The data analysis revealed several critical insights that directly supported the purpose and function of the dashboard. The analysis showed that churn rates were relatively consistent across age groups, with a significant spike in churn for customers aged 60 and above in the Kaggle (JB Link Telco) dataset. This insight is crucial for stakeholders looking to implement targeted retention strategies for older customers. The data also indicated that churn rates were similar across genders, suggesting that gender is not a strong predictor of churn so stakeholders can focus on other factors for their retention strategies. The dashboard revealed that average monthly charges were lower in the Kaggle dataset compared to WGU but unlike the WGU data, we see a sharp spike in monthly charge once the users reach about 65 years old in the Kaggle dataset. This could suggest a targeted strategy may need more focus for this older demographic. Lastly, tenure was consistent across both datasets. These findings support the idea that while pricing strategies may differ, customer loyalty, as measured by tenure, remains similar.

C6: Analysis Limitations

While the data analysis provided valuable insights, there are some limitations to consider. Differences in data structure and format between the internal and external datasets required extensive cleaning and transformation. The analysis was also limited to specific variables (age, gender, zip code, monthly charges, and tenure). Other potentially influential factors, such as customer service interactions or product offerings, were not included, which could limit the comprehensiveness of the analysis. The datasets analyzed come from different sources (WGU vs. Kaggle), and there may be inherent biases in the data that were not accounted for. For example, differences in customer demographics, service offerings, or market conditions could impact the results. Lastly, some metrics were analyzed at a high level (e.g., age groups), which may obscure more granular trends within sub-groups of customers. Despite these limitations, the analysis successfully identified key patterns and provided actionable insights to stakeholders that can guide them in improving customer retention.

D: Web Sources

1. "Import Data from CSV into Postgre" Retrieved from https://hasura.io/docs/latest/schema/postgres/postgres-guides/import-data-from-csv/
2. "Join Your Data" Retrieved from https://help.tableau.com/current/pro/desktop/en-us/joining_tables.htm

E: Sources

1. "WGU_D210_Task1_Justin_Huynh.pdf" Retrieved from Task 1 from D210
2. "D210.twb" Retrieved from Task 1 from D210
3. "JB Link Telco Customer Churn" Retrieved from https://www.kaggle.com/datasets/johnflag/jb-link-telco-customer-churn