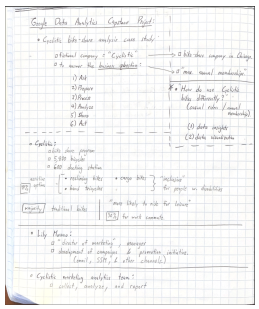
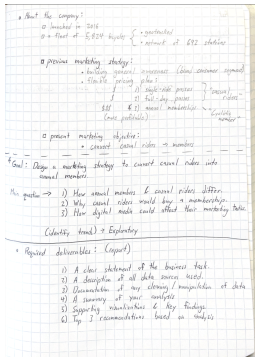
Pre-work:

- Familiarize myself with the context of the problem at hand.
    a. Documented in a notebook.
        i. Company, stakeholders, previous/current state, future objective, main goal, main approach, and required deliverables.
        ii. 
        iii. 

- Start a guideline that oversees the activities and scope of the project.
    a. Google doc: “Cyclistic Capstone Project (Pre-work)”
    b. Link:

- Create a log of the progress that occur during this project.
    a. Google doc: “Cyclistic Capstone Project (Log)”

- In the guidelines, draft the scope of the project.
    a. 4 main components:
        i. Data Analysis Approach:
            1. (Ask, Prepare, Process, Analyze, Share, and Act)
        ii. Marketing Research Approach:
            1. a
        iii. Technical Approach:
            1. Track everything; Measure everything.
        iv. Presentational Approach:
            1. Setting up and organizing the insights in a presentable manner.

Data Processing Log:

1. Created a folder “CapstoneProject_Cyclistic_Data” to house project files.
2. Created 3 subfolders within the “CapstoneProject_Cyclistic_Data” folder to house, ZIP files, CSV files, and XLS files, named, “zip_files”, “csv_files”, “xls_files” respectively.
3. Downloaded the ZIP files of roughly the last 12 months of trip data (Jan 2024 through Dec 2025), and moved them into the “zip_files” folder.
4. Unzip the 12 downloaded ZIP files resulting in 12 CSV files.
5. Moved the 12 CSV files into the “csv_files” folder.
6. Launch Excel,
    a. select the Data tab, Get Data (Power Query), Text/CSV, browse, select the 12 csv files.
7. Save and name the workbook “Cyclistic_12m_of_tripdata_template”
8. Duplicate the workbook and name the duplicate “Cyclistic_12m_of_tripdata_(edited)”
9. Open “Cyclistic_12m_of_tripdata_(edited)”
10. Rename each spread sheet in (yyyy-mm) format.
11. “Convert to range” the tables in each spreadsheet.
12. Due to data privacy constraints, personally identifiable information will be removed, fields: “ride_id, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng” in each spreadsheet.
13. Remove table accents on all spreadsheets.
    a. Highlight all columns, -> Format cells -> Font -> Color: Black -> Fill -> Background color: No Color -> Border -> None -> ok.
14. Checked for duplicates, with unique function:
    a. (data - remove duplicates)
    b. No duplicate data.
15. Rename column “member_casual” to “MembershipType”.
16. Move column “MembershipType” to column A.
17. Rename, “rideable_type” to “RideType”
18. Rename, “started_at” to “StartTimeStamp”
19. Rename, “ended_at” to “ReturnTimeStamp”
20. create 4 new columns for both “StartTimeStamp” and “ReturnTimeStamp” called, “StartTime, StartDate, StartDay, StartMonth” and “ReturnTime, ReturnDate, ReturnDay, ReturnMonth” respectively.
21. For “StartTimeStamp” and “ReturnTimeStamp” change custom format from “m/d/yy h:mm” to “mm-dd-yyyy hh:mm:ss” on a 24 hour clock for each spreadsheet.
22. For the columns “StartTime” and “ReturnTime”, extract the time from “StartTimeStamp” and “ReturnTimeStamp” respectively using =MOD(cell_ref,1).
    a. For “StartTime” and “ReturnTime”, change custom format from “m/d/yy h:mm” to “h:mm:ss”.
23. For the columns “StartDate” and “ReturnDate”, extract the time from “StartTimeStamp” and “ReturnTimeStamp” respectively using =INT(cell_ref).

    a. For “StartTime” and “ReturnTime”, change custom format from “m/d/yy h:mm” to “mm-dd-yyyy”.
24. For the columns “StartDay” and “ReturnDay”, extract the time from “StartTimeStamp” and “ReturnTimeStamp” respectively using =TEXT(cell_ref, “dddd”).
25. For the columns “StartMonth” and “ReturnMonth”, extract the time from “StartTimeStamp” and “ReturnTimeStamp” respectively using =TEXT(cell_ref, “mmmm”).
26. Create a ride duration column.
    a. Create a column named “RideLength”
        i. Subtract “ReturnTimeStamp” and “StartTimeStamp”
        ii. Format column as custom -> [h]:mm:ss
            1. To allow greater than 24 hr times.
        iii. Copy and paste value
            1. Note: There are negative time values in this data.
                a. 11-2024: 43 instances
                b. 05-2024: 76 instances
                c. 04-2024: 58 instances
                d. 03-2024: 25 instances
                e. 02-2024: 5 instances
                f. 01-2024: 20 instances
                    i. They were removed:
                        1. Sort ride length-> smallest to largest-> delete
27. Create a column RideLengthSeconds that converts RideLength to Seconds:
    a. = (HOUR(cell_with_time) * 3600) + (MINUTE(cell_with_time) * 60) + SECOND(cell_with_time)
28. Copy and paste the value for all columns.

Data Analysis Log:

1. Copy each spreadsheet into its unique workbook, titled “mm-yyyy_data”
2. Open each spreadsheet & convert data to a table in the spreadsheet “mm-yyyy”

3. Create a pivot table and in a new workbook called “mm-yyyy_analysis”.

    i. % of Riders by Membership Type:
        1. Rows: MembershipType
        2. Columns: Values
        3. Values: count of MembershipType, count of MembershipType
            a. Show data of one count of MembershipType to “% of Grand Total”
            b. Rename both as “# of Riders” and “% of Riders” respectively

    ii. Bike Type by Membership Type:

1. Rows: MembershipType
2. Columns: RideType
3. Values: count of MembershipType
    a. Show data of the count of MembershipType to “% of Grand Total”
    b. Rename as “BikeType by Membership”

iii. Start Time by Membership Type:
1. Rows: Hours(StartTime)
    a. Group -> Hours
2. Columns: MembershipType
3. Values: count of MembershipType
    a. Rename as “# of Riders”

iv. Average Ride Duration by Membership Type:
1. Rows: (none)
2. Columns: MembershipType
3. Values: RideLength
    a. Change to Average RideLength
    b. Show as time hh:mm:ss

v. Descriptive Statistics of Ride Length:
1. Rows: MembershipType
2. Columns: Values
3. Values: RideLength (4x)
    a. Display Average, Min, Max, and StdDev
    b. Name appropriately
    c. Design -> Grand Total -> off for Row & Col
4. Find Median of RideLengthSeconds off of 2-way ANOVA and convert for MembershipType.
    a. =TEXT(A1/86400, "hh:mm:ss")

vi. # of Riders by Ride Duration:
1. PivotTable
    a. Rows: RideLength
    b. Columns: MembershipType
    c. Values: Count of MembershipType

vii. # of Rides by Day of Week:
1. Rows: StartDay
2. Columns: Values
3. Values: Count of MembershipType 2x
    a. Change one to % of the Grand Total

viii. Average Ride Duration By Day of Week:

1. Rows: StartDay
2. Columns: (none)
3. Values: Average of Ridelength
    a. Change to time format hh:mm:ss

ix. Chi-square of Independence (MembershipType & RideType):
1. H0: RideType is independent of MembershipType
2. H1: RideType is related to MembershipType
3. Pivot Table:
    a. Rows: MembershipType
    b. Columns: RideType
    c. Values: Count of MembershipType

x. Chi-square of independence (MembershipType |to| frequency of weekday & weekend uses)
1. H0: MembershipType is independent of weekly bike usage
2. H1: MembershipType is related to weekly bike usage
3. Pivot Table:
    a. Rows: MembershipType
    b. Columns: StartDay (group by weekday and weekend)
    c. Values: Count of MembershipType

xi. Chi-square of independence (RideType |to| frequency of weekday & weekend uses)
1. H0: RideType is independent of weekly bike usage
2. H1: RideType is related to weekly bike usage
3. Pivot Table:
    a. Rows: RideType
    b. Columns: StartDay (group by weekday and weekend)
    c. Values: Count of MembershipType

xii. Chi-square GOF (MembershipType) [*Note could indicate certain months to target the marketing campaign in]

xiii. Chi-square GOF (RideType) [indicate they are not the same frequency (is a difference)]

xiv. 2-way ANOVA (MembershipType & BikeType |to| RideLengthSeconds.) w replication
1. Create new sheet: “ANOVA_worksheet”
2. Pivot Table (Original Data):
    a. Row: MembershipType
    b. Column: RideType
    c. Values: Count of MembershipType

        i. To determine how many samples in each group
3. Copy and paste columns:
    a. MembershipType, RideType, and RideLengthSeconds
    b. Clear format
4. Sort columns:
    a. MembershipType A-Z, RideType A-Z
5. Conditional Format:
    a. Highlight Cells: (specific text) (not really necessary)
        i. Casual
        ii. Member
        iii. Electric_bike
        iv. Classic_bike
6. Add Filter to columns
7. Copy and paste each group:
    a. Column: MembershipType
    b. Rows: RideType
8. Create random column:
    a. =rand()
    b. Copy paste values
    c. Sort by random column for each group (smallest to largest)
        i. 10,000 sample
            1. Using =SEQUENCE(10000,1) for quicker data selection
9. Data analysis -> Anova: Two-Factor With Replication
    a. Sample (RideType)
    b. Columns (MembershipType)
    c. Post Hoc Analysis:
        i. Tukey's HSD test
            1. 4 to 6 groups:
                a. (casual classic_bike, casual electric_bike, member classic_bike, member electric_bike; casual electric_scooter, member electric_scooter)
            2. 3!=6 to 5!=15 comparisons: (N groups)
                a. (N*(N-1))/2
            3. Q:
                a. # of groups
                b. Degree of freedom
                    i. 4 groups => 3.633
            4. Critical Range:
                a. Q*SQRT(WithinMS/replications)
            5. Absolute Mean Difference:
                a. Compare AMD to CR
                    i. > sig

ii. < not sig
iii. Conditional formate (“not”) -> [red]

6. Direction of difference:
    a. If $\bar{x}_1 - \bar{x}_2 > 0$; $\bar{x}_1 > \bar{x}_2$
    b. If $\bar{x}_1 - \bar{x}_2 < 0$; $\bar{x}_1 < \bar{x}_2$

- SpreadSheet completed:
    - ☑ ~~09-2024~~
    - ☑ ~~08-2024~~
    - ☑ ~~07-2024~~
    - ☑ ~~06-2024~~
    - ☑ ~~05-2024~~
    - ☑ ~~04-2024~~
    - ☑ ~~03-2024~~
    - ☑ ~~02-2024~~
    - ☑ ~~01-2024~~
    - ☑ ~~12-2023~~
    - ☑ ~~11-2023~~
    - ☑ ~~10-2023~~

Data Visualization Log:

1. Charts:

HEX *#FFC20A,* **R***255***G***194***B***10;* “casual”

HEX *#0C7BDC,* **R***12***G***123***B***220;* “member”

HEX *#4B0092,* **R***75***G***0***B***146;* “general”

HEX *#332288,* **R***51***G***34***B***136;* “classic_bike”

HEX *#88CCEE,* **R***136***G***204***B***238;* “electric_bike”

HEX *#882255,* **R***136***G***34***B***85;* “electric_scooter”

a. % of Riders by Membership Type:
    i. PivotChart -> Pie chart
        1. HEX #FFC20A
        2. HEX #0C7BDC

b. Bike Type by Membership Type:
    i. PivotChart -> Bar chart
        1. HEX #332288
        2. HEX #88CCEE
        3. HEX #882255

c. Start Time by Membership Type:
    i. PivotChart -> Bar chart (horizontal)
        1. HEX #FFC20A
        2. HEX #0C7BDC
        3. Change time category in reverse order.

d. Average Ride Duration by Membership Type:
    i. PivotChart -> Bar chart
        1. HEX #FFC20A
        2. HEX #0C7BDC

e. # of Riders by Ride Duration:
    i. PivotChart -> Bar chart
        1. HEX #4B0092

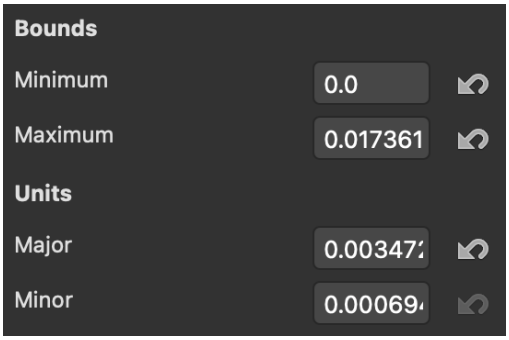
f. # of Rides by Day of Week:
    i. PivotChart -> Bar chart
        1. HEX #4B0092

g. Average Ride Duration By Day of Week:
    i. PivotChart -> Bar chart
        1. HEX #FFC20A
        2. HEX #0C7BDC

2. Dashboard:
    a. Merge rows 1-8 from columns C onwards and Center vertically.
    b. Insert, “Cyclistic Monthly Report: month year”, font size 65
    c. Insert Cyclistic Logo
    d. Insert graphs by importance:
        i. (Format-> Arrange-> Align-> Snap to Grid)
        ii. Group: “% of Riders by Membership, Avg. Ride Duration by Membership, & Start time by Membership”.
            1. % of Riders by Membership:(hxl) 3.94” x 4.51”
            2. Avg. Ride Duration by Membership: 3.94” x 4.51”
            3. Start time by Membership: 8.1” x 9.55”
        iii. Group: “# of Riders by Day of Week, Average Ride Duration by Day of Week, # of Rider by Ride Duration”.
            1. # of Riders by Day of Week: 4.17” x 4.51”
            2. Average Ride Duration by Day of Week: 4.17” x 4.51”
            3. # of Rider by Ride Duration: 4.86” x 14.44”
        iv. Bike type by membership type.
            1. Bike type by membership type: 4.17” x 4.97”
    e. Add Slicer:
        i. MembershipType
        ii. RideType
            1. Slicer-> Snap to Grid
            2. Rclick-> report connections-> check boxes:
                a. PivotTable #
                b. Chart with % on axis, needs a compatible range for all slicer options. (max 100%)

Lessons Learned:

- Determine which column fields will be useful in finding insights about the business task beforehand to save on computing power.
- Sort the data by reasonable column fields beforehand for better organization (primary key).
- Excel handles dates and times using a system in which dates are serial numbers and times are fractional values. For example, June 1, 2000 12:00 PM is represented in Excel as the number 36678.5, where 36678 is the date (June 1, 2000) and .5 is the time (12:00 PM). In other words, the time value in a "datetime" is the decimal.
    - Time conversion in excel from min to fraction for graph axis.

- (ex: 0-25min, by 5min & 1min) 

- A #SPILL! Error can occur if a function in one cell creates a range of data, like =SEQUENCE function, when converting to a table. To fix this, copy and paste value.
- Split spreadsheets into separate workbooks early to minimize computation power necessary to complete tasks.
- When testing out specific analysis, test on a smaller dataset for quicker results.
- 2-way ANOVA in Excel:
    - “The main effects are the portion of the relationship between an independent variable and the dependent variable that does not change based on the values of the other variables in the model.”
    - “Interaction effects indicate that another variable influences the relationship between an independent and dependent variable.”
    - Sample maximum of around 10,000.
- You can find PivotTable name in Rightclick-> pivotable options
- Trendline is removed when slicer is used.
- Slicers disconnect graphs with different data sources.
- Make sure graph range fits all slicer variations of the graph.