Penelope Newcomb, Emily Isaacson, Salma Tagnaouti,
Dhruvit Patel, and Anjolie Mavani
Pamela Duffy
INST 327 0101

# Team Project Final Submission

# Introduction

## Problem Domain and Motivation

We decided to focus on the the metro stations in Maryland, specifically the ones we found needed to prioritized for improvements. This topic is extremely important to all of us as we have access to a metro station very easily and with the construction of the Purple Line, the metro will be on campus. Initially, to get a better understanding, we were only going to focus on Montgomery County and Prince George's County, however during the process, we decided to add Baltimore County, so our data could be more representative. Also, our analysis was focused around how we can move forward and identify stations that needed to be improved, but specifically we were focused on the accessibility each station provided. If the accessibility could be increased, riders' experience would overall increase as well, leading to more customers and moe funding towards public transportation efforts.

## Database Introduction

Our first table is the Bikes Table. This includes a primary Bike_ID key along with several columns to give viewers a better understanding of what the bike parking and locking situation looks like in each station. Related is our Station Access table which identifies stations by their Station_ID along with their Access_ID to assign each station a pedestrian and bike access score. Next, we have our Station_Type table which identifies each station as one of a few different station types including an anchor station which is a station that links multiple types of public transportation (such as Metro buses and the LightRail). The Station_Lines table links each rail line to a station. We had originally combined this information with the Rail_Lines table however we wanted to avoid unary relations in our ERD. Rail_Lines uses the primary key of Rail_Type to differentiate between different lines of public transportation as the MARC train as well as other public transportations services run several different lines in differing directions. Our TOD table allows us to identify the type of neighborhood a station is located in. Each TOD_ID is linked to a different environment including Downtown or Village Center.

Next, we have the Amenities table which contains information about the various amenities at different transit stations. These amenities include shelter, benches, public restroom, and public phone and they were each assigned an Amenity ID. We also have a Station Amenities table, which acts as a composite table for Amenities. This table assigns the different stations an Amenity ID. This table also specifies if the amenities exist or not. In addition, the Construction Permits table contains data on the types of permits available including residential new construction, nonresidential new construction, and mixed use new construction. These types of permits were each assigned a Construction Permit ID.

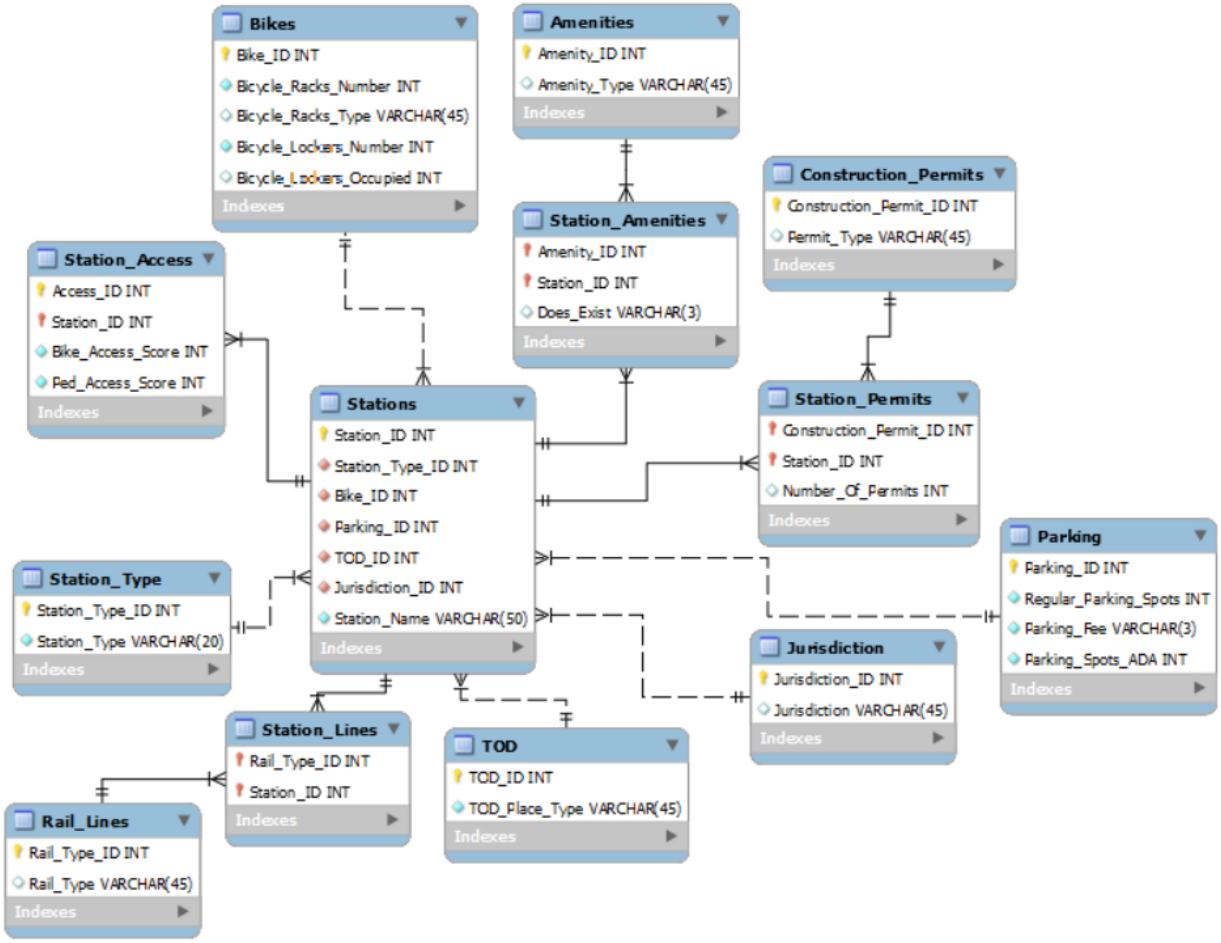
The Station Permits table acts as a composite table that tells us the different types of permits for each transit station, as well as the number of permits for each station using Construction_Permits_ID to assign each station (Station_ID) a type of permit. Moreover, the Parking table contains data on the availability of regular and ADA parking spaces as well as if there is a fee or not. We used Parking_ID as a primary key in order to identify each of the

station's parking availability. Lastly, the jurisdiction table provides information on the three different jurisdictions that we are analyzing for this project and assigning them a number using Jurisdiciton_ID. Finally we have our primary linking table, Stations. This table allows us to link several foreign key IDs to get a better understanding of the general statistics surrounding a specific station. We have included Station_ID, Station_Type_ID, Bike_ID, Parking_ID, TOD_ID, as well a non numeric value which gives viewers the chance to see which ID belongs to which specific station name.

Our first query used the stations amenities table, station amenities, and the station access table to examine how a station's access score correlated to how many amenities it has. Our second query used the stations, parking, and station access table to examine how a station's access score correlated to how many parking spots it has. Our third query uses the stations, station type, station access, and jurisdiction tables to examine the difference in access scores between jurisdictions. Our fourth query uses the stations, TOD, station permits, and construction permits tables to examine the difference in number and type of construction permits by station type. Our fifth query uses the stations, station lines, rail lines, and station access table to examine whether a station serving more rail lines correlates to a higher access score. Our sixth query uses the stations, station access, station permits, and construction permits tables to examine whether stations with low access scores have many construction permits, and therefore may be scheduled for improvement. And finally, our seventh query uses the stations, station type, and station access table to examine the relationship between station type and access score.

# Database Description

## Logical Design




## Physical Database

Our database has thirteen tables. We have eighteen individual tables giving us eighteen individual Station_IDs. The Station_ID has allowed us to link each station to a specific access score, bike and parking situations, as well as the amenities and permits that each station has. Other tables act as points of information for tables that have foreign keys and and act as linking tables between the main Stations table and the tables with primary keys other than Station_ID.

## Sample Data

All of our comes from stations highlighted in the MTA Transit database. We selected eighteen stations from the counties of: Montgomery, Prince Georges, and Baltimore City. Most of the information that is used to evaluate and measure these specific stations can be found in the Stations table which lists a variety of linking IDs to allow for those viewing the table to get a better understanding of what a station's general capabilities are. For more information and to

view the data we used, please look at our backup SQL file as well as the text file that demonstrates how we utilized that data.

**Views / Queries**

| View name | Has a JOIN | Uses filtering | Uses aggregation | Linking table and both source tables | Subquery |
|---|---|---|---|---|---|
| amenity_acess_scores | X | | X | X | |
| station_parking_capacity | X | X | X | | |
| jurisdiction_access | X | | X | | |
| rail_line_access_scores | X | X | | X | |
| station_improvement | X | X | X | X | X |
| place_type_construction_ permits | X | X | | X | |
| station_type_access | X | | X | | |

**Changes from Original Design**

For our ERD, we decided to add a “Rail_Lines" table to “Station_Lines” because when we included Rail_Type_ID as a primary key in the Stations_Lines table, we got a unary relationship between Rail_Type_ID and Station_ID. In addition, we realized that we did not have the minimum sample records in our database as well as there being no construction permit records present in the Montgomery County stations and the Prince George’s County stations. In order to solve this problem, we decided to add another jurisdiction, Baltimore City County, so we could have more sample records and construction permits present in our database.

Furthermore, our original plan was to make sure that most columns were not null because we wanted to provide complete information to our audience. However, we soon realized that it was not possible given that some of the data in the database was null, and we had to include it regardless. In order to include this data, we had to check off “not null” for most of the columns, which allowed us to include some sample records that have some null and not null columns. We also had an issue with adding “Amenity_ID " to the “Parks” an’d “Bikes” table because it created a unary relationship. We decided to get rid of “Amenity_ID” in these tables in order to solve this issue.

By getting a better understanding of how tables are linked in the entire database, we were able to figure out how every table was connected to each other within the database and come up

with solutions to our unary relationship errors. Lastly, we decided to update the VARCHAR() datatype in some columns such as “Station_Name” to 50 characters in order to allow us to import our data successfully.

## Database Ethics Considerations

### Diversity, Equity, and Inclusion Considerations

After reviewing our database and making revisions as needed, we believe that our database seems likely to be an inclusive information resource in its domain. The main goal of our database is to look at how accessible the stations are in Prince George’s County, Montgomery County, and Baltimore City County and look at ways to improve station accessibility in these counties. One potential bias is how the data is collected since the database may not be representative of current station conditions which can lead to an incomplete understanding of accessibility.

In addition, there could be some data quality bias; some information may be missing such as the TOD place type and station type which can lead to an incomplete analysis of accessibility. In order to minimize the bias, we need to carefully consider the information presented in each of the columns by assessing which information should be null or not null. We also need to ensure that we are using a good variety of SQL queries in order to analyze accessibility better. In order to provide complete information to our audiences, we have decided to check off important columns that should be not null including the station type, access scores, access type, and tod place type in order to have a more mature, inclusive, and complete design.

### Data Privacy, Fair Use, Other Ethical Considerations

After reviewing our database and solidifying our plan we can now more confidently identify what possible obstacles we could run into with our data. There does not seem to be any pressing concerns in terms of privacy and other ethical considerations, as the ERD does take into account all the possible demographics. The only problem that can possibly arise will be the misuse of the data but the chances of that is very small and otherwise there should not be any other issues. Even if that situation arises it is doubtfully will make an impactful difference on the final database.

## Reflection and Next Steps

### Lessons Learned

Over the course of this semester we learned a few lessons that we will take into consideration when completing other database administration assignments and queries. First is to always cross reference our data and the questions we are trying to answer and solve. We had decided to focus on Montgomery County and Prince Georges County when we were first brainstorming with our database. However, since we were looking to focus heavily on the amount of permits within our views and queries we would need to add more stations later on in our process. By double checking that the stations we had selected had a data point for each

column, we will be able to save time in the future that would otherwise be wasted reuploading and importing files to populate our database.

Along the same lines as double checking the amount and validity of our data, we learned the lesson that it is better to ‘measure twice and to cut once’. By reviewing each step and table we continued or importing, we would have not had to reupload CSV files as many times and would reduce the amount of errors we had in both forward engineering and populating our database. Through the making of mind maps or other organizational structures to track our progress, there would have been more of an incentive to double check our work before ‘officially’ crossing it off of our to-do list. Ultimately, through frequent communication via Discord and iMessages we were able to relieve ourselves of any extremely large mistakes that would require an extensive amount of reworking. However, through more frequent in person meetings we may have been able to get a better understanding of what each person was working on as well as their work speed to allocate tasks more accordingly in the future.

**Potential Future Work**

If we had more time for our database there are definitely some aspects we might have implemented differently, but for the most part the final database focused on everything we wanted to target and served the purpose of our project. One change we might have made would have been to add more stations to cover an even wider area or as mentioned above, finding all of data points of each station earlier in the process, so populating the database would have been easier. There is always room for improvement, however with our limited resources and time we were able to create a database that targeted the information we wanted to explore.