CSI4142 Fundamentals of Data Science

Project Phase 3 - OLAP Queries and BI Dashboard

April 5, 2024

Shannon Noah - 300163898

Serena Iyoha - 300187757

Tina Trinh - 300175427

GitHub Repo Link to Project: https://github.com/tinatrinh8/Data-Science-Gentrification-Study
(Scripts to execute the SQL queries are found in "Phase 3 OLAP Queries" Folder)

Part A. OLAP queries

Part 1. Standard OLAP operations

a. Drill down and roll up

Query 1: Drilling down the date dimension to get the number of permits by quarter

Query | Query History

```
SELECT d."Year", d."Quarter", COUNT(*) AS "TotalPermitsPerQuarter"
FROM "BuildingPermitsDimension" p
JOIN "DateDimension" d ON p."Application_Date_Key" = d."Date_Key"
GROUP BY d."Year", d."Quarter"
ORDER BY d."Year", d."Quarter";
```

Data Output

	Year integer	Quarter integer	TotalPermitsPerQuarter bigint
1	1999	4	5
2	2000	1	9
3	2000	2	9
4	2000	3	10
5	2000	4	6
6	2001	1	8
7	2001	2	8
8	2001	3	11
9	2001	4	7

Total rows: 98 of 98 | Query complete 00:00:00.169

Query 2: Drilling down to get the number of permits by street name, rather than per Ward

Query | Query History

```
SELECT
    (p."Street_Name" || ' ' || p."Street_Type" || ' ' ||
     COALESCE(p."Street_Direction", '')) AS "FullStreetName",
    COUNT(*) AS "TotalPermits"
FROM
    "BuildingPermitsDimension" p
GROUP BY
    "FullStreetName"
ORDER BY
    "FullStreetName";
```

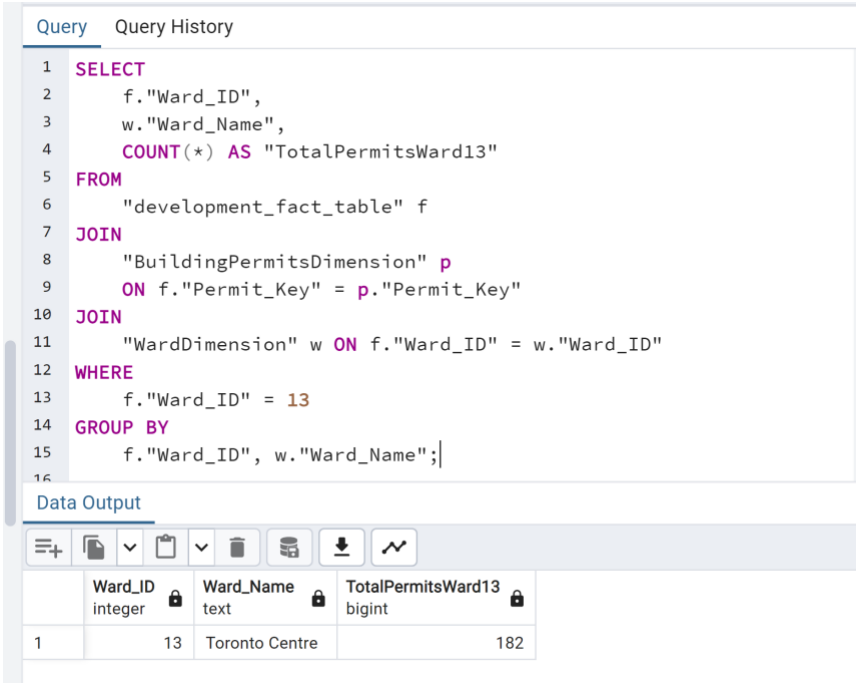
Data Output

	FullStreetName text	TotalPermits bigint
1	ABBEYWOOD TRL	2
2	ABBOTSFORD RD	1
3	ABELL ST	1
4	ABILENE DR	3
5	ABINGER CRES	1
6	ABITIBI AVE	7

Total rows: 1000 of 2351 | Query complete 00:00:00.192

b. Slice

Query 3: Slicing to get the total number of building permits in the 13th Ward



c. Dice

Query 4: Dicing to form a sub-cube, which has dimensions 'Ward_ID' and 'Year'

Query Query History

```
SELECT
    f."Ward_ID",
    d."Year",
    COUNT(*) AS "TotalPermits"
FROM
    "development_fact_table" f
JOIN
    "BuildingPermitsDimension" p ON f."Permit_Key" = p."Permit_Key"
JOIN
    "DateDimension" d ON p."Application_Date_Key" = d."Date_Key"
GROUP BY
    f."Ward_ID",
```

Data Output

	Ward_ID integer	Year integer	TotalPermits bigint
1	0	1999	1
2	0	2004	11
3	0	2008	1
4	0	2010	3

Query 5: Dices the data to show annual totals for demolitions and replacements of affordable, mid-range, and high-end rental homes.

Query Query History

```
SELECT
    d."Year",
    COUNT(*) AS "TotalDemolitions",
    SUM(dd."Affordable Rental Homes for Demolition") AS "TotalAffordableDemolitions",
    SUM(dd."Mid-Range Rental Homes for Demolition") AS "TotalMidRangeDemolitions",
    SUM(dd."High-End Rental Homes for Demolition") AS "TotalHighEndDemolitions",
    SUM(dd."Affordable Rental Homes Replaced") AS "TotalAffordableReplacements",
    SUM(dd."Mid-Range Rental Homes Replaced") AS "TotalMidRangeReplacements",
    SUM(dd."High-End Rental Homes Replaced") AS "TotalHighEndReplacements"
FROM
    "DemolitionDimension" dd
JOIN
    "DateDimension" d ON dd."Approval_Date_Key" = d."Date_Key"
GROUP BY
    d."Year"
ORDER BY
    d."Year";
```

	Year integer	TotalDemolitions bigint	TotalAffordableDemolitions numeric	TotalMidRangeDemolitions numeric	TotalHighEndDemolitions numeric	TotalAffordableF numeric
1	2017	11	184	101	13	
2	2018	6	150	174	8	
3	2019	17	323	397	23	
4	2020	12	152	39	39	
5	2021	9	295	31	26	
6	2022	23	519	268	80	
7	2023	24	1047	511	49	

*Note that "TotalDemolitions" represents the count of demolition events, while the other columns, like “TotalAffordableDemolitions” represent the count of individual housing units affected by those demolitions.

d. Combining OLAP operations

Query 6: Dices the data by year and occupation type to calculate the total population and the average ratio of individuals without certifications for each occupation.

```
SELECT
    e."Employment" AS "OccupationType",
    w."Year",
    SUM(w."Population") AS "TotalPopulationForOccupation",
    AVG(w."No_Cert_Ratio") AS "AverageNoCertRatio"
FROM
    "EmploymentDimension" e
JOIN
    "ward_profile_fact_table" w ON e."Employment_Key" = w."Dimension_Key"
GROUP BY
    e."Employment",
    w."Year"
ORDER BY
    w."Year",
    e."Employment";
```

Data Output

	OccupationType text	Year bigint	TotalPopulationForOccupation double precision	AverageNoCertR double precision
1	All occupations	2016	4590990	0.16683161000
2	Business, finance and administration occupations	2016	784215	0.16653528937
3	Health occupations	2016	602570	0.1663692546
4	Management occupations	2016	769655	0.16461633587
5	Natural and applied sciences and related occupations	2016	563000	0.1649047392
6	Natural resources, agriculture and related production occupations	2016	431335	0.16469697073
7	Occupations in art, culture, recreation and sport	2016	488380	0.1673460498(
8	Occupations in education, law and social, community and government servic...	2016	857275	0.16436058888
9	Occupations in manufacturing and utilities	2016	532800	0.16637349357
10	Sales and service occupations	2016	1181390	0.16643540356

From 2016 to 2021, there seems to be a greater percentage of the population who have a diploma or certificate (16% to 14% no certificate ratio) indicating an increase in the number of individuals with diplomas or certifications over time in these employment groups.

Query 7: Analyzing how shelter costs and the percentage of income spent on shelter vary over time and by household type (Owned vs. Rented)

Query | Query History

```
SELECT
    sd."Household_Type",
    wp."Year",
    AVG(sd."Average_Monthly_Shelter_Costs") AS "AvgShelterCosts",
    AVG(sd."Percent_Spending_30_Percent_Or_More_On_Shelter")
    AS "AvgPercentSpendingMoreThan30Pct",
    SUM(wp."Population") AS "TotalPopulation"
FROM
    "ShelterDimension" sd
JOIN
    "ward_profile_fact_table" wp ON sd."Shelter_Key" = wp."Dimension_Key"
GROUP BY
    sd."Household_Type",
    wp."Year"
ORDER BY
    wp."Year",
    sd."Household_Type";
```

Data Output

	Household_Type text	Year bigint	AvgShelterCosts numeric	AvgPercentSpendingMoreThan30Pct double precision	TotalPopulation double precision
1	Owner	2016	1677.7307692307692308	27.35	1174155
2	Tenant	2016	1209.4615384615384615	46.64615384615385	1051645
3	Owner	2021	2027.4615384615384615	25.423076923076923	1205830
4	Tenant	2021	1513.0769230769230769	38.84615384615385	1115940

Query 8: Compares the change in population for each occupation type between 2016 and 2021 by summing up the populations across all wards

Query | Query History

```
WITH Population2016 AS (
    SELECT
        e."Employment" AS "OccupationType",
        SUM(w."Population") AS "Population2016"
    FROM
        "EmploymentDimension" e
    JOIN
        "ward_profile_fact_table" w ON e."Employment_Key" = w."Dimension_Key"
    WHERE
        w."Year" = 2016
    GROUP BY
        e."Employment"
),
Population2021 AS (
    SELECT
        e."Employment" AS "OccupationType",
        SUM(w."Population") AS "Population2021"
    FROM
        "EmploymentDimension" e
    JOIN
        "ward_profile_fact_table" w ON e."Employment_Key" = w."Dimension_Key"
    WHERE
```

```
        w."Year" = 2021
    GROUP BY
        e."Employment"
)
SELECT
    p2016."OccupationType",
    p2016."Population2016",
    p2021."Population2021",
    (p2021."Population2021" - p2016."Population2016") AS "ChangeInPopulation"
FROM
    Population2016 p2016
JOIN
    Population2021 p2021 ON p2016."OccupationType" = p2021."OccupationType"
ORDER BY
    "ChangeInPopulation" DESC;
```

Data Output

	OccupationType text	Population2016 double precision	Population2021 double precision	ChangeInPopulation double precision
1	Sales and service occupations	1181390	2205210	1023820
2	Natural and applied sciences and related occupations	563000	543270	-19730
3	Business, finance and administration occupations	784215	709350	-74865
4	Trades, transport and equipment operators and related occupations	698845	559460	-139385
5	Occupations in art, culture, recreation and sport	488380	344435	-143945
6	Health occupations	602570	339270	-263300
7	Occupations in manufacturing and utilities	532800	203735	-329065
8	Occupations in education, law and social, community and government servic...	857275	479710	-377565
9	Natural resources, agriculture and related production occupations	431335	28020	-403315
10	Management occupations	769655	310585	-459070
11	All occupations	4590990	3117675	-1473315

Total rows: 11 of 11 Query complete 00:00:00.096

Query 9: Slice and dice to get the permits for new houses where the estimated construction cost is over 10000000

Query Query History

```
1  SELECT
2      bp."Permit_Key",
3      bp."Permit_Type",
4      df."Est_Const_Cost",
5      df."Ward_ID"
6  FROM
7      "BuildingPermitsDimension" bp
8  JOIN
9      development_fact_table df ON bp."Permit_Key" = df."Permit_Key"
10 WHERE
11     df."Est_Const_Cost" > 10000000
12     AND bp."Permit_Type" = 'New Houses'
13
```

Data Output

	Permit_Key bigint	Permit_Type text	Est_Const_Cost integer	Ward_ID integer
1	574	New Houses	23000000	23
2	3337	New Houses	20000000	11
3	4489	New Houses	12000000	11
4	4512	New Houses	11000000	11
5	4646	New Houses	15000000	11
6	5922	New Houses	61013000	2

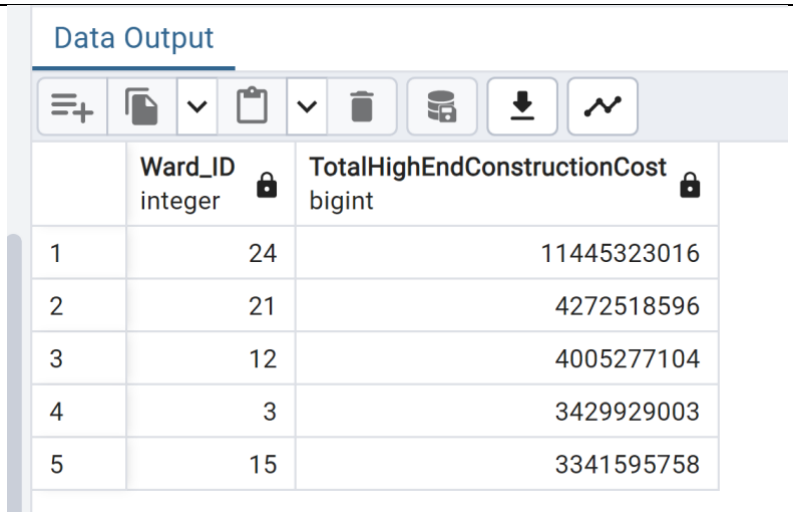
Part 2. Explorative operation

a. Iceberg

Query 10: Returns the top five wards with the highest total estimated construction costs above $50,000, suggesting areas with the most high-investment construction activity

Query | Query History

```
SELECT
    dft."Ward_ID",
    SUM(dft."Est_Const_Cost") AS "TotalHighEndConstructionCost"
FROM
    "development_fact_table" dft
WHERE
    dft."Est_Const_Cost" > 50000
GROUP BY
    dft."Ward_ID"
ORDER BY
    "TotalHighEndConstructionCost" DESC
LIMIT 5;
```



	Ward_ID integer	TotalHighEndConstructionCost bigint
1	24	11445323016
2	21	4272518596
3	12	4005277104
4	3	3429929003
5	15	3341595758

b. Widowing

Query 11: Ranks wards by the average project duration for construction projects, displaying the ranks per ward for each of the last five years.

Query Query History

```sql
SELECT
    dft."Ward_ID",
    dd."Year",
    AVG(dft."Application_to_Issuance_Duration") AS "AverageProjectDuration",
    RANK() OVER (PARTITION BY dd."Year" ORDER BY AVG(dft."Application_to_Issuance_Duration") DESC)
FROM
    "development_fact_table" dft
JOIN
    "DateDimension" dd ON dft."Application_Date_Key" = dd."Date_Key"
WHERE
    dd."Year" BETWEEN EXTRACT(YEAR FROM CURRENT_DATE) - 5 AND EXTRACT(YEAR FROM CURRENT_DATE)
GROUP BY
    dft."Ward_ID",
    dd."Year"
ORDER BY
    dd."Year" DESC, "AverageProjectDuration" DESC;
```

Data Output

	Ward_ID integer	Year integer	AverageProjectDuration numeric	DurationRank bigint
1	13	2024	56.0000000000000000	1
2	3	2024	45.0000000000000000	2
3	4	2024	40.0000000000000000	3
4	20	2024	40.0000000000000000	3
5	1	2024	36.6000000000000000	5
6	12	2024	35.0000000000000000	6
7	17	2024	33.5000000000000000	7
8	9	2024	32.0000000000000000	8
9	0	2024	30.5000000000000000	9

c. Using the window clause

Query 12: Compares the number of building permit issued per ward to that of the previous and next years

Query Query History

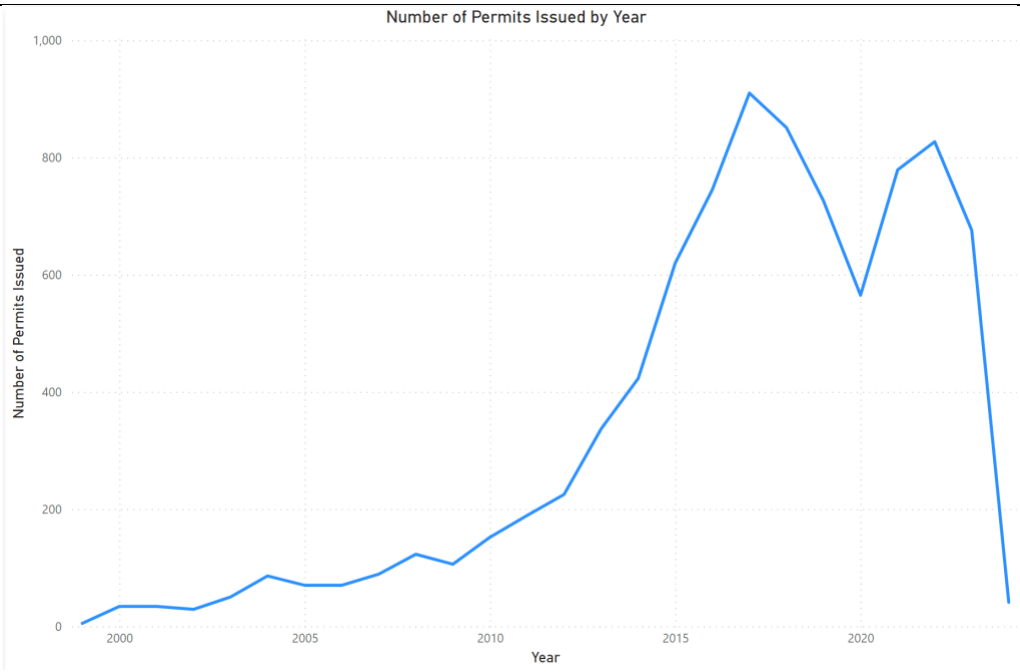
```sql
SELECT
    dft."Ward_ID",
    dd."Year",
    COUNT(dft."Permit_Key") AS PermitsIssued,
    LAG(COUNT(dft."Permit_Key"), 1) OVER (PARTITION BY dft."Ward_ID" ORDER BY dd."Year") AS Previ
    LEAD(COUNT(dft."Permit_Key"), 1) OVER (PARTITION BY dft."Ward_ID" ORDER BY dd."Year") AS Next
FROM
    "development_fact_table" dft
JOIN
    "DateDimension" dd ON dft."Issued_Date_Key" = dd."Date_Key"
GROUP BY
    dft."Ward_ID",
    dd."Year";
```

Data Output

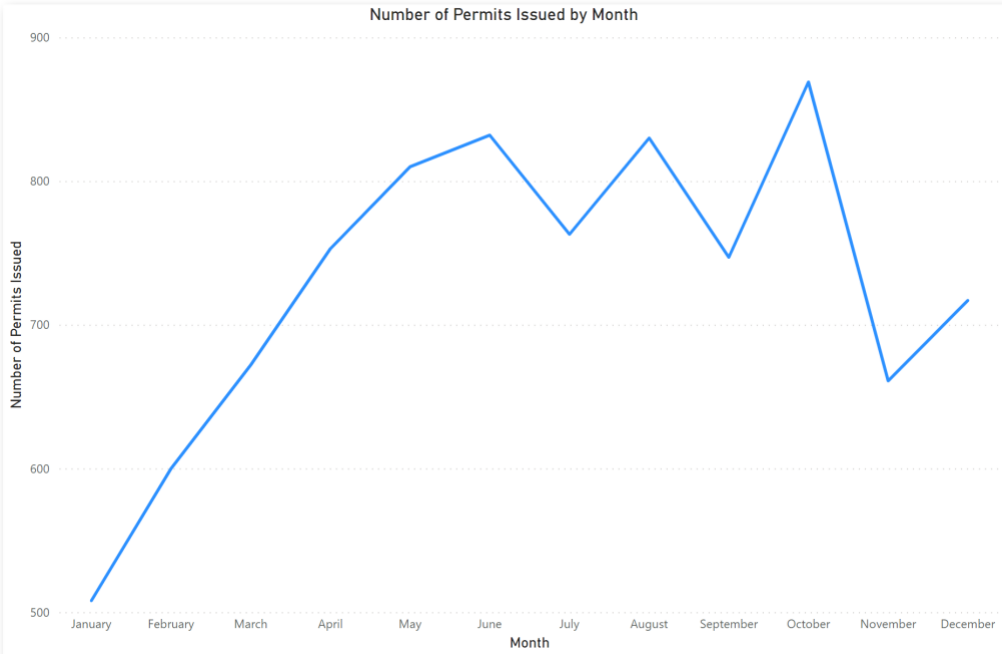
	Ward_ID integer	Year integer	permitsissued bigint	previousyearpermits bigint	nextyearpermits bigint
1	0	2000	1	[null]	11
2	0	2007	11	1	1
3	0	2008	1	11	2
4	0	2010	2	1	2
5	0	2011	2	2	2
6	0	2012	2	2	2
7	0	2013	2	2	2
8	0	2014	2	2	7
9	0	2015	7	2	3
10	0	2016	3	7	10
11	0	2017	10	3	9
12	0	2018	9	10	5
13	0	2019	5	9	7

Part B. BI dashboard and Information Visualization

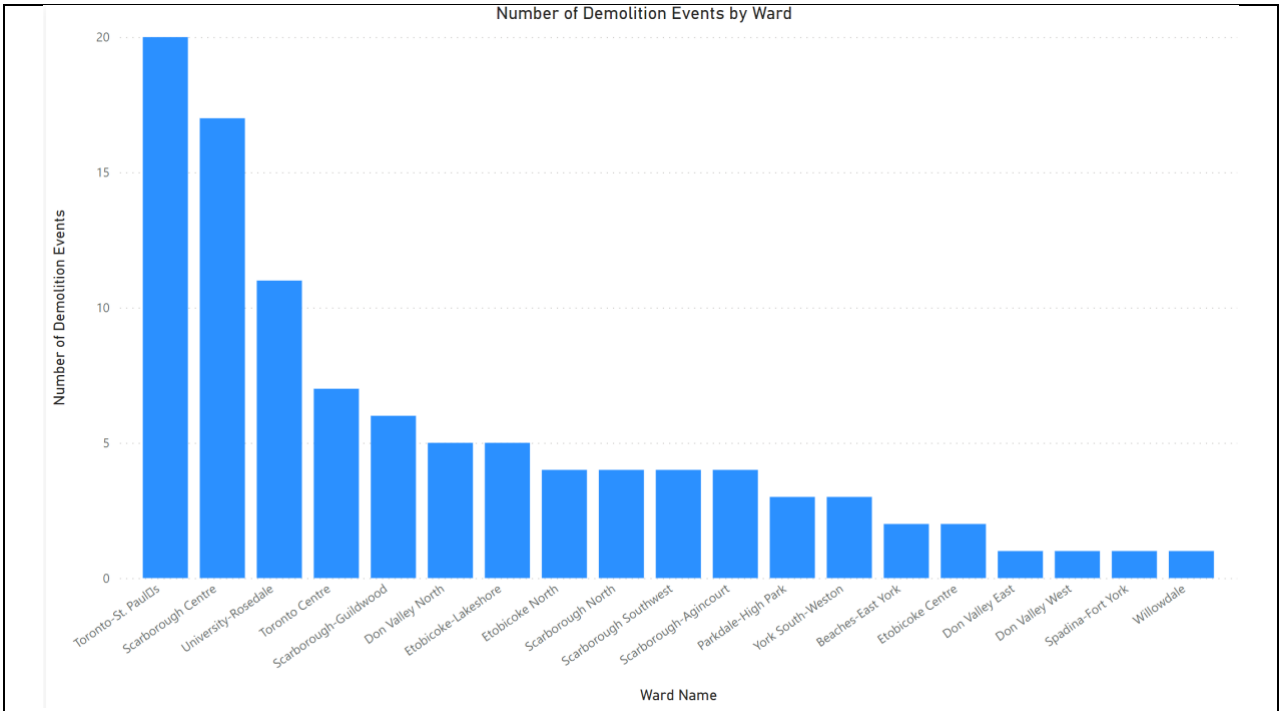
Visualization 1: Compares the number of permits issued over the years



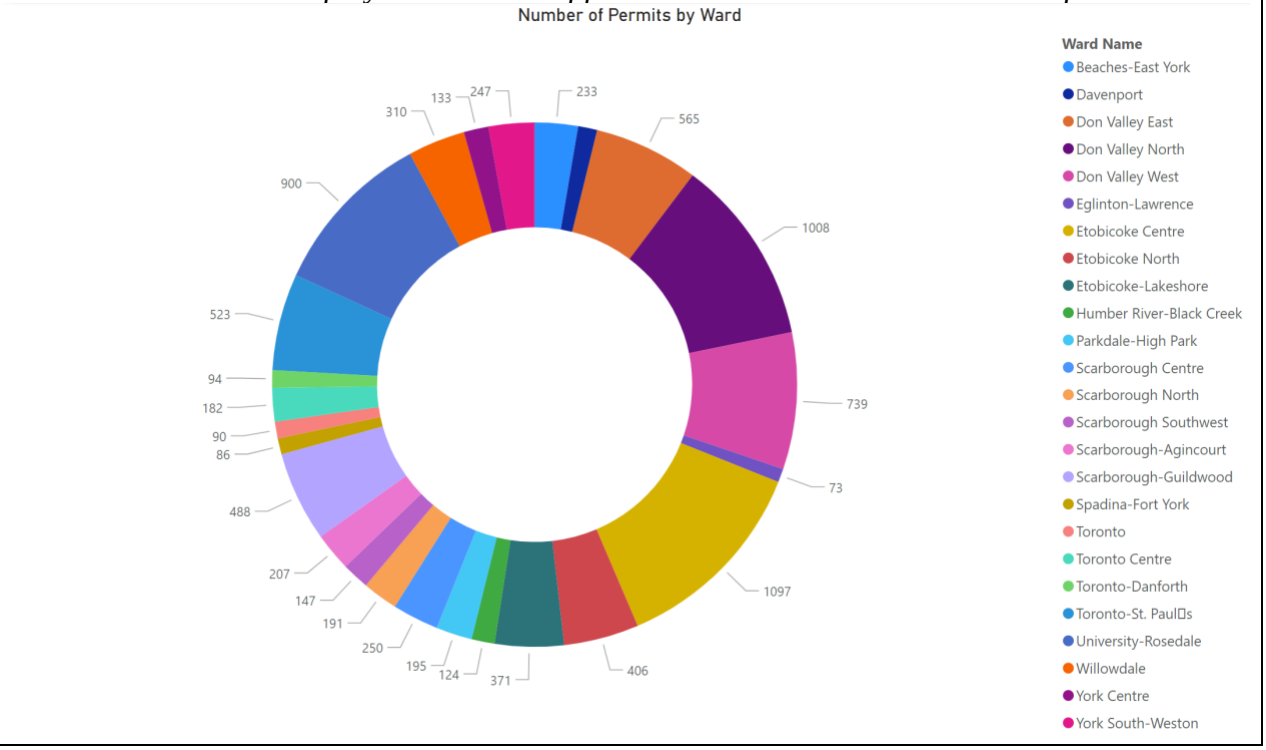
Ability to drill down to months using the date dimension:



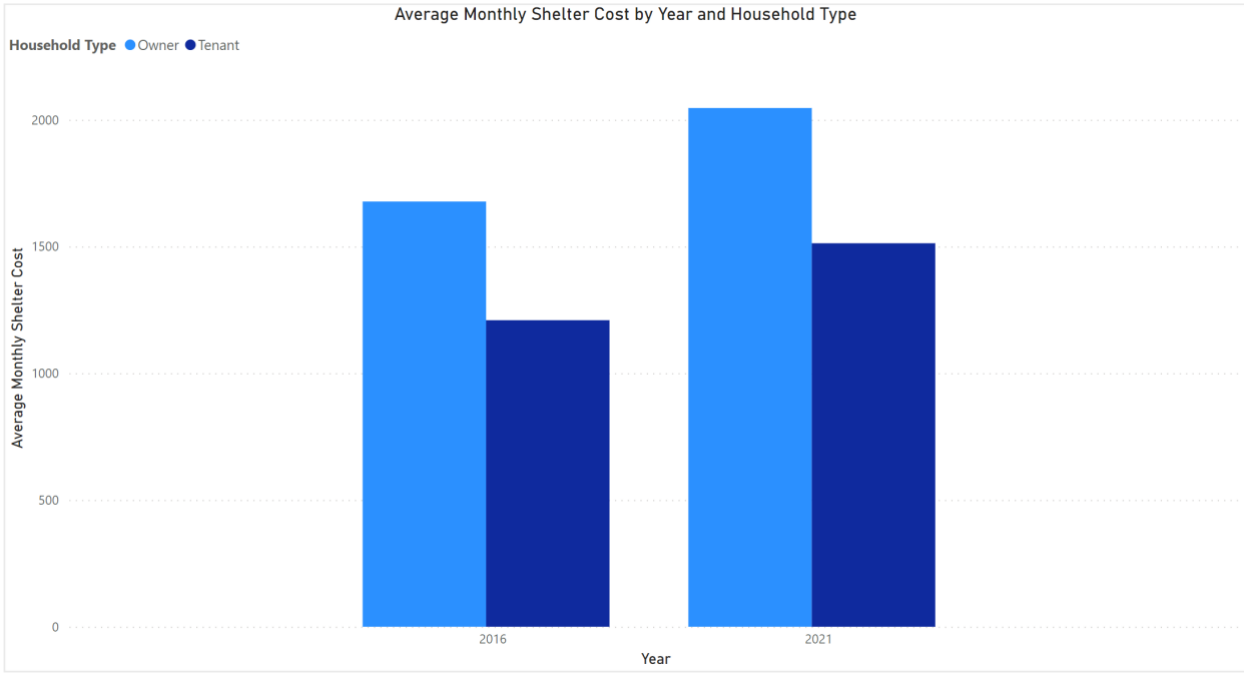
Visualization 2: Examines the number of demolition events occurring within each ward



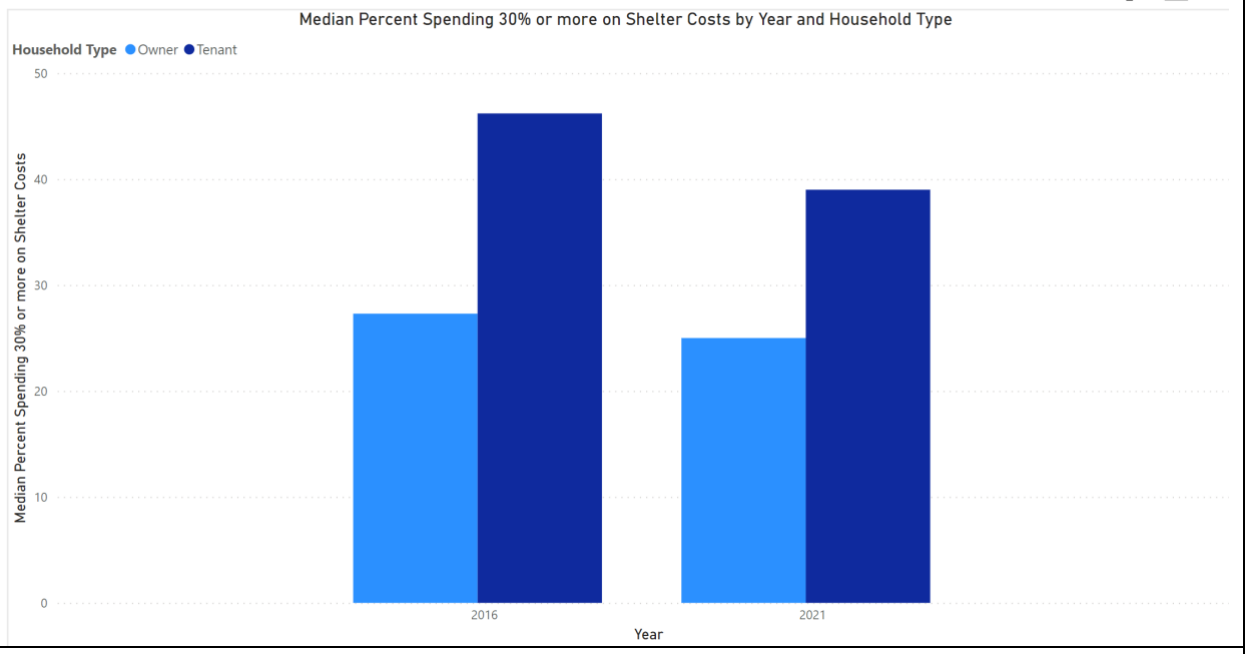
Visualization 3: Displays the number of permits issued within each ward in a pie chart



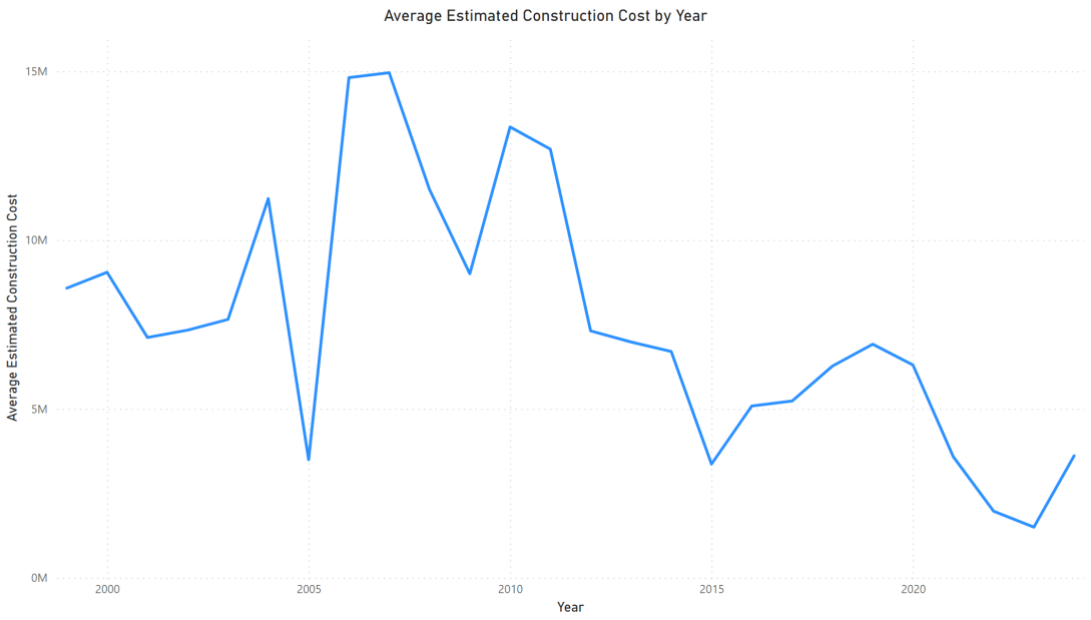
Visualization 4: Compares the average monthly shelter costs of Owners vs. Tenants



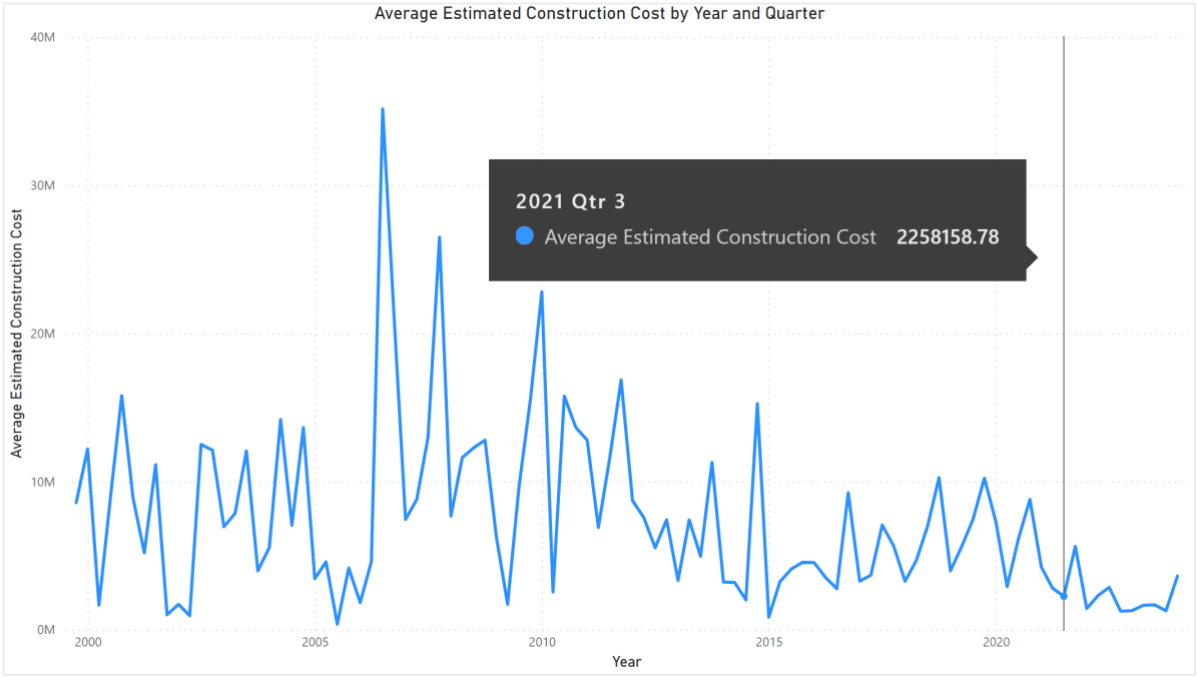
Visualization 5: Compares the median percentage of population spending 30% or more of income on shelter costs



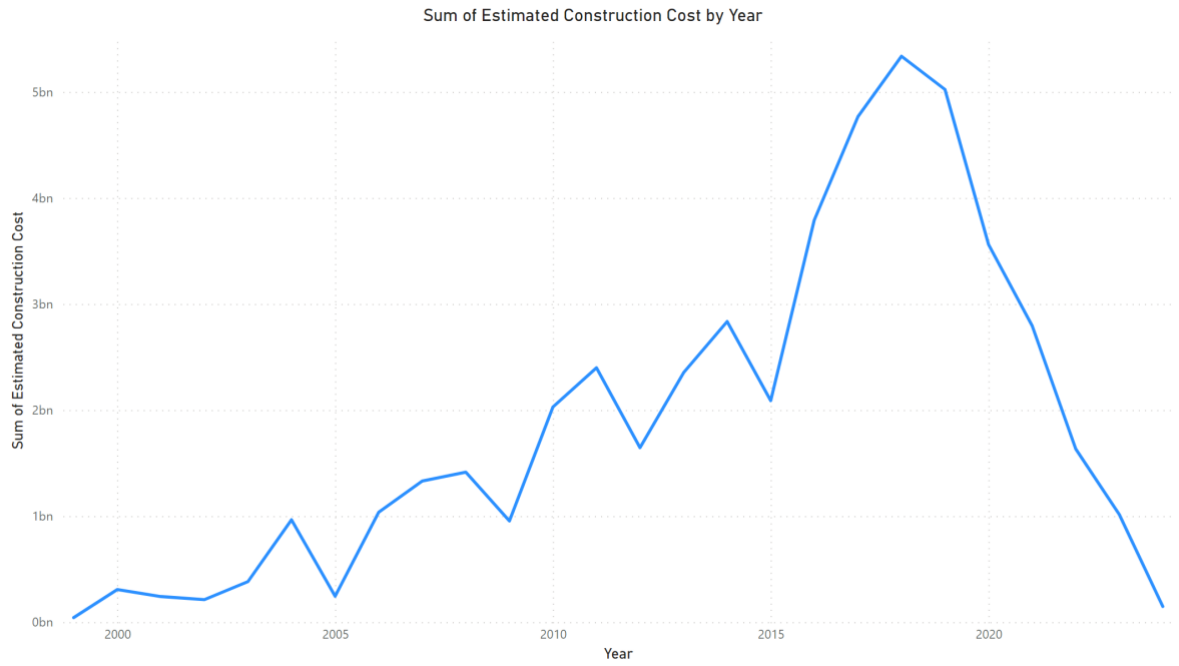
Visualization 6: Displays the trends in the average estimated construction cost for all wards over the years



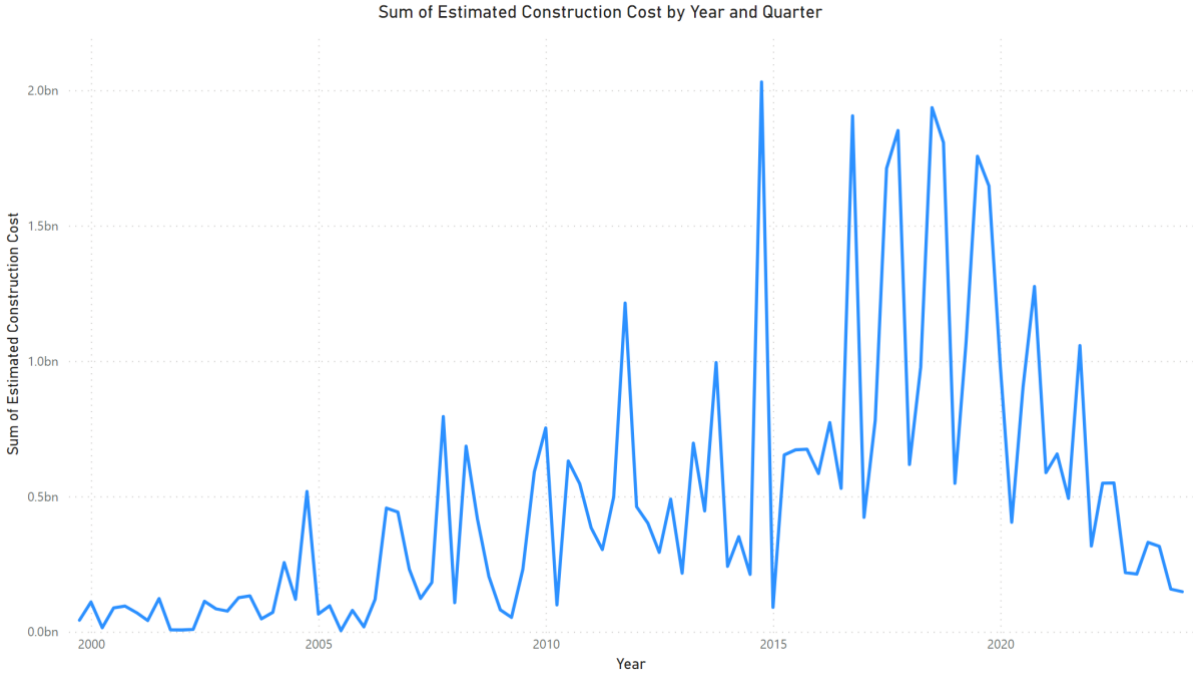
Ability to drill-down to show quarter with date dimension:



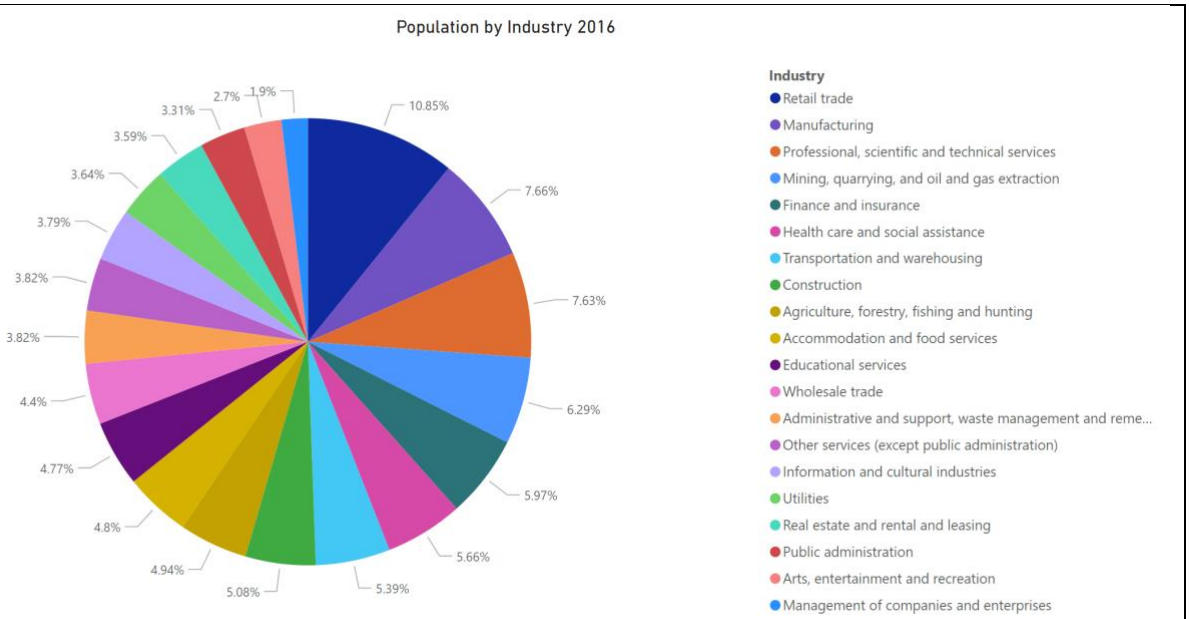
Visualization 7: Displays the trends in the sum of estimated construction cost for all wards over the years



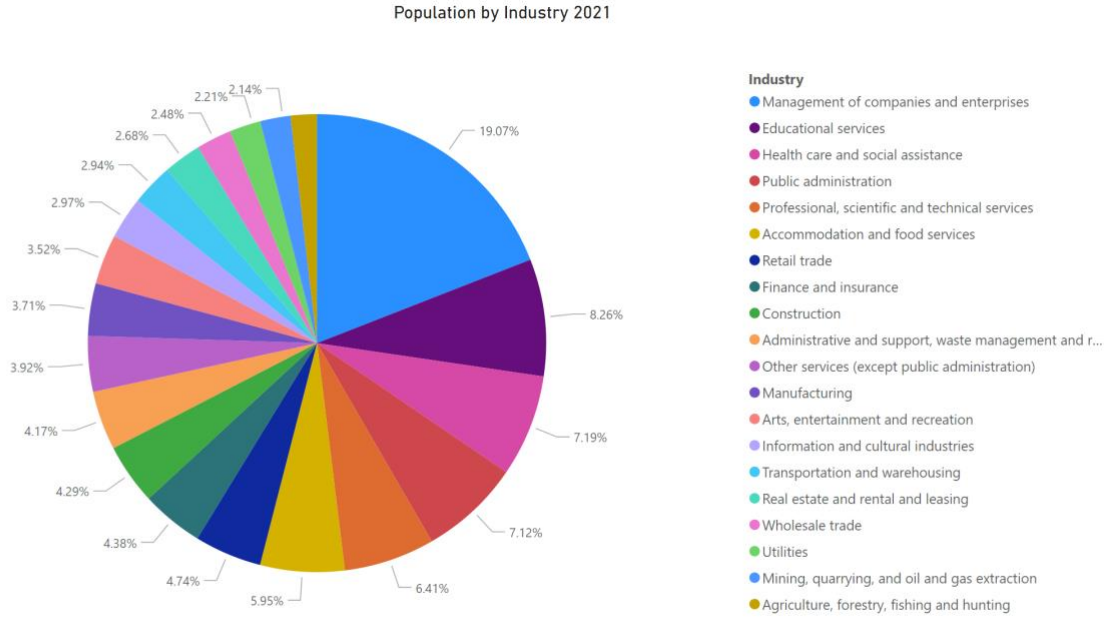
Ability to drill-down to show quarter with date dimension:



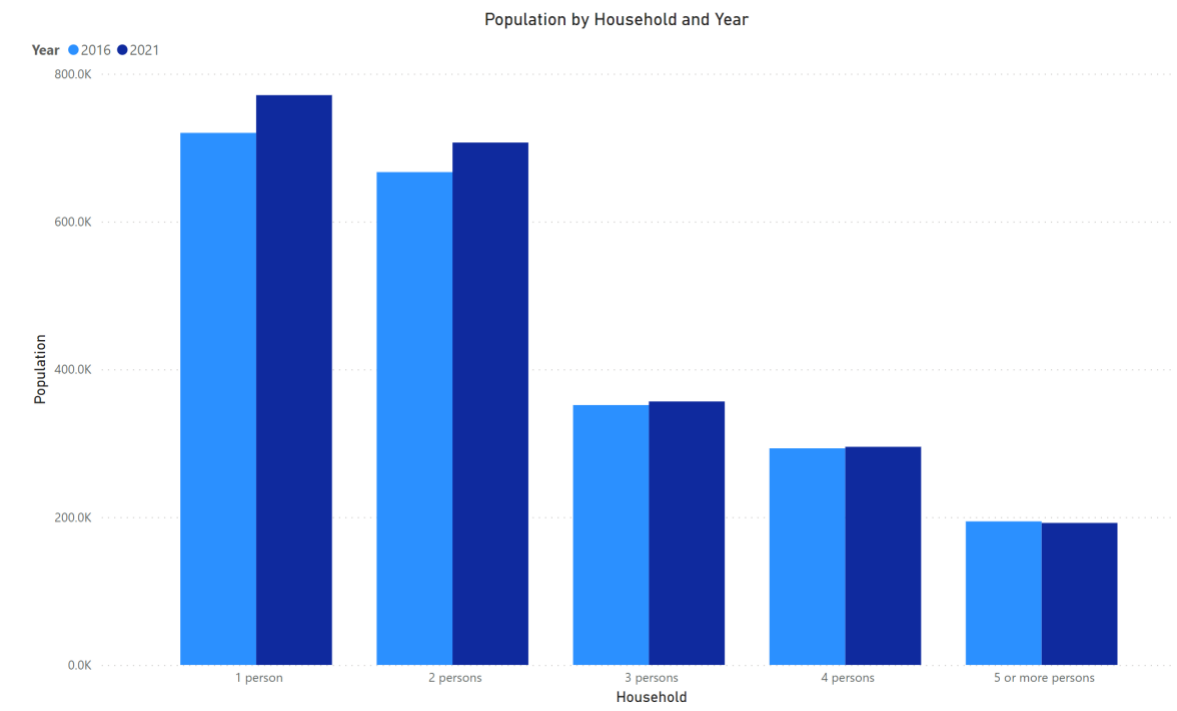
Visualization 9: Displays the percentage of the population participating in each employment industry type for the year 2016



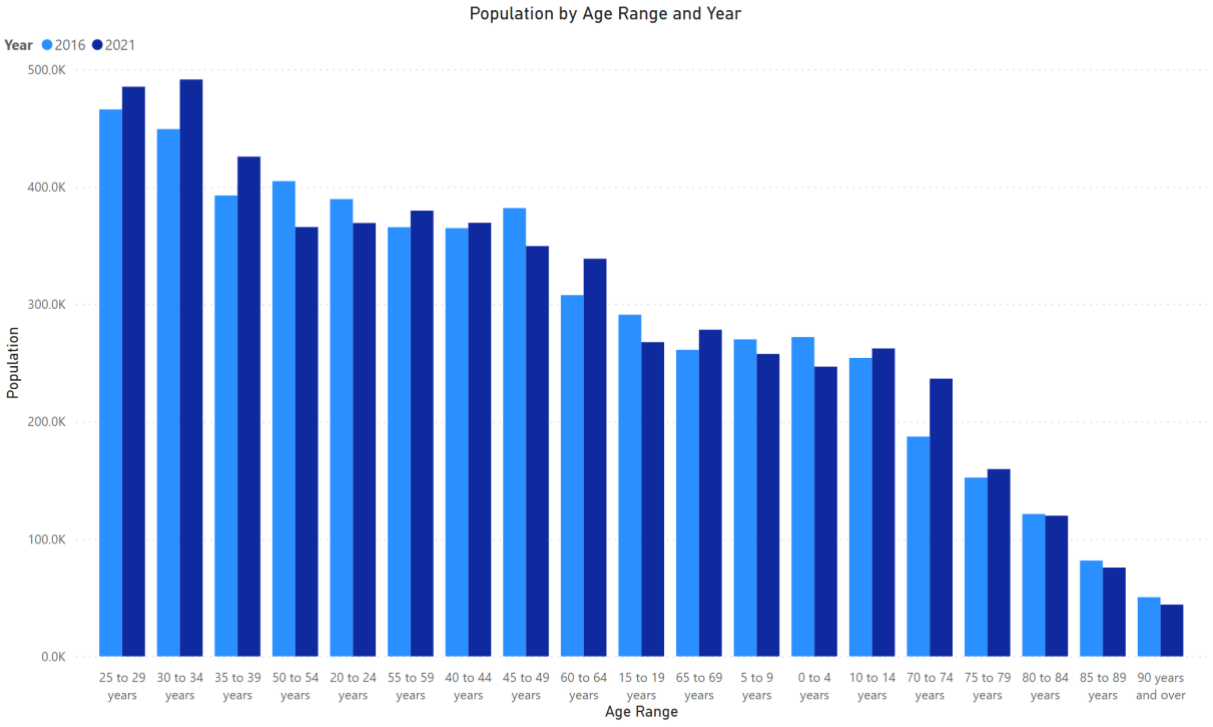
Visualization 10: Displays the percentage of the population participating in each employment industry type for the year 2021



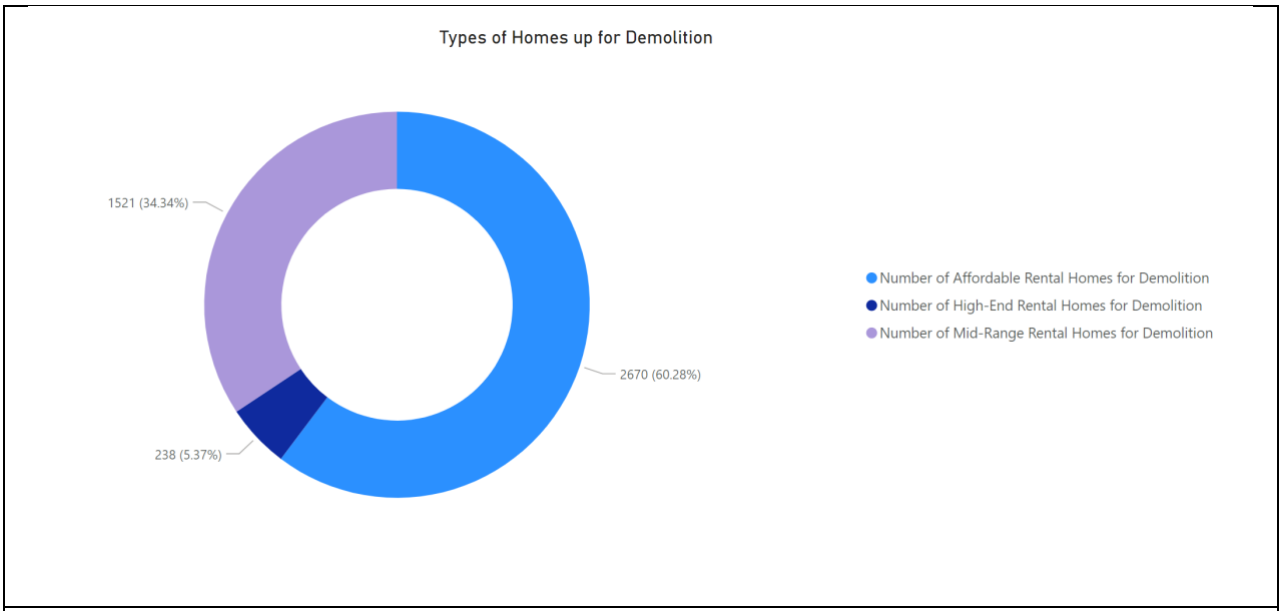
Visualization 11: Comparing the change in population for household size from 2016 to 2021



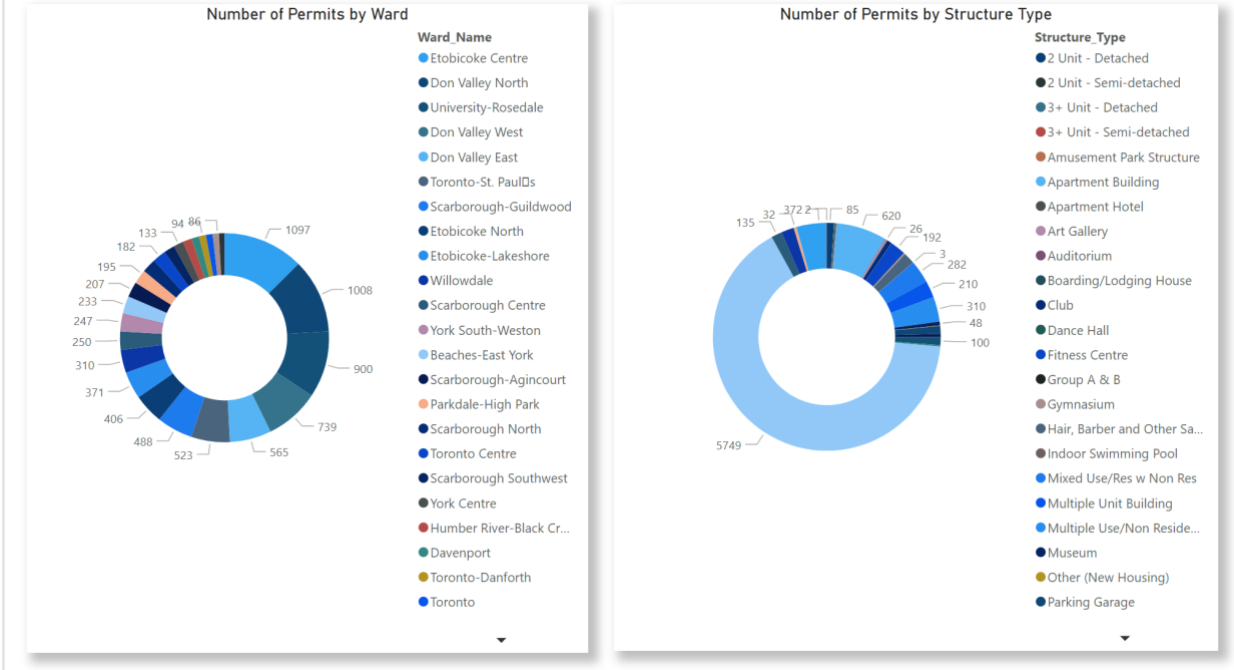
Visualization 12: Comparing the change in population for age brackets from 2016 to 2021



Visualization 13: Displaying the number of total demolitions for each type of home

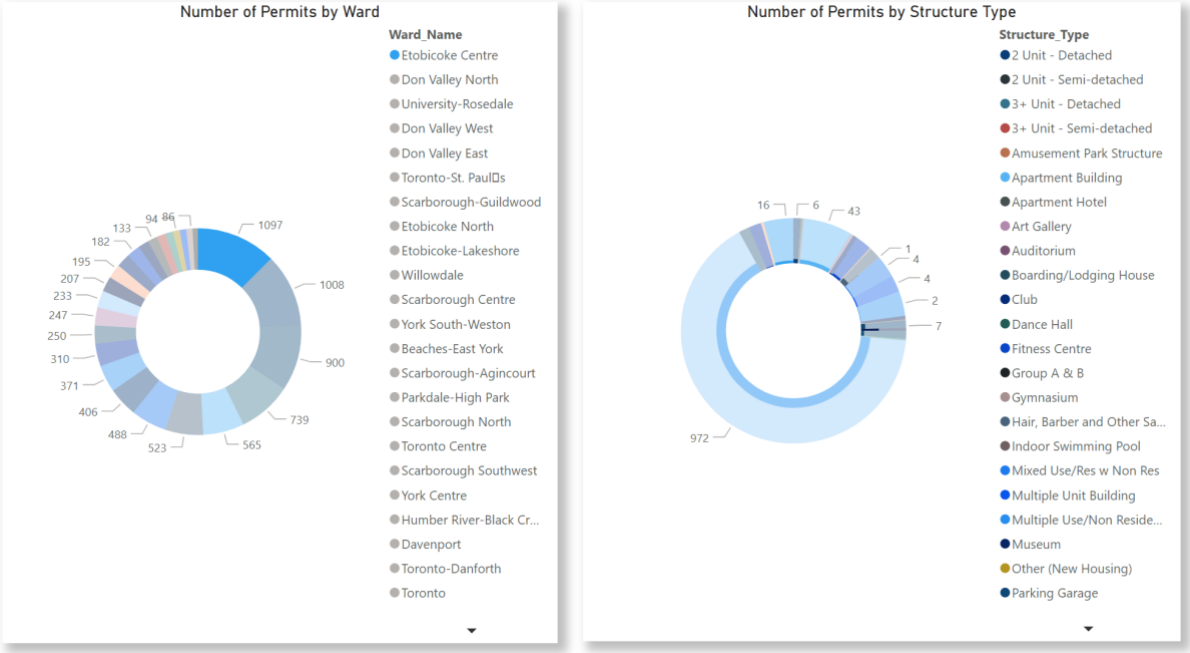


Visualizations 14 & 15: Displaying the number of permits by ward and structure type

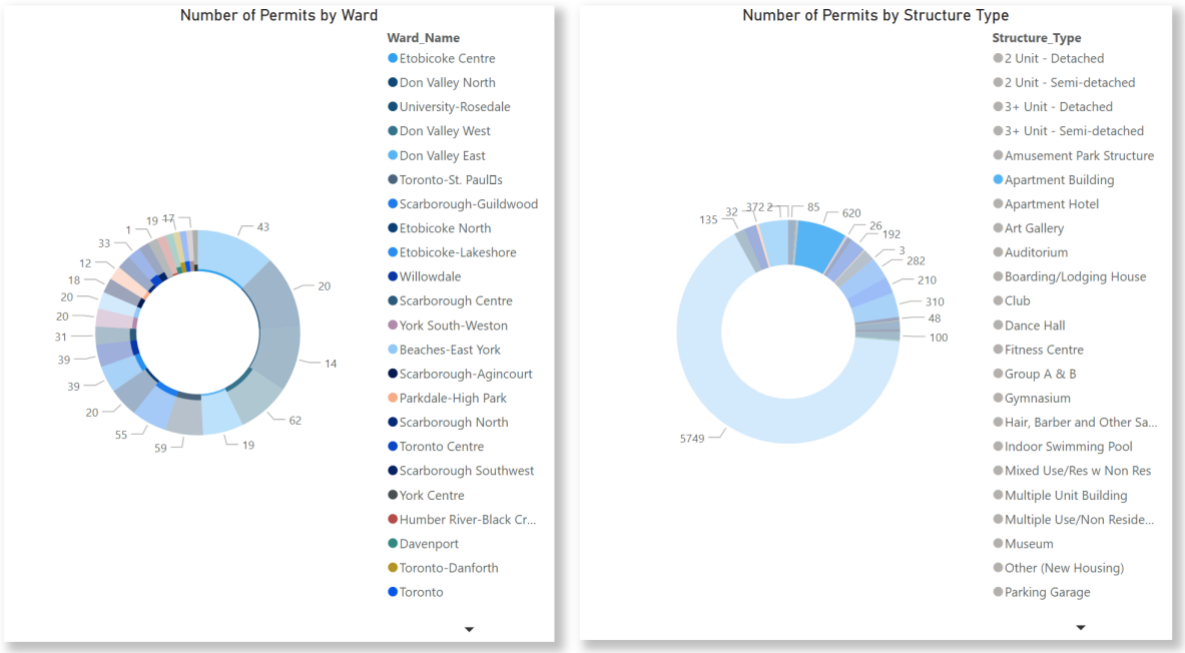


Slicing and Dicing

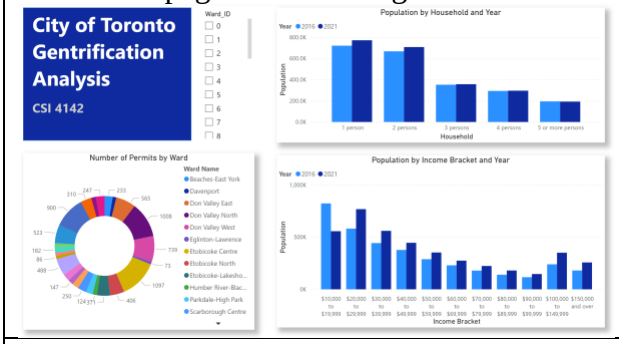
By selecting Etobicoke Center as the ward, the user can see the breakdown of the structure types of the permits associated with that ward:



By selecting structure type “Apartment Building”, the user can see in which wards the permits of this structure type originated from:

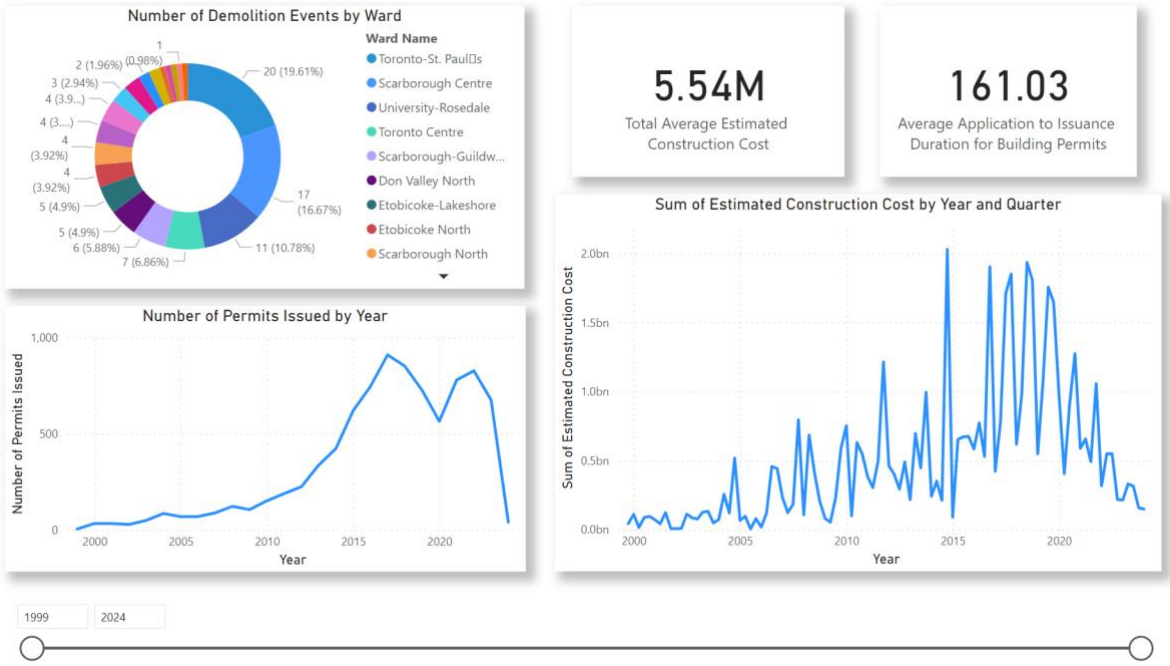


Dashboard page before slicing:

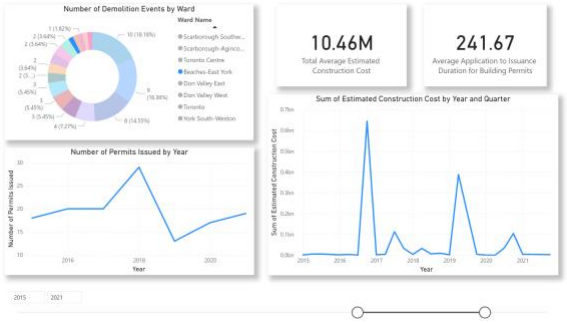


After slicing for ward id = 3

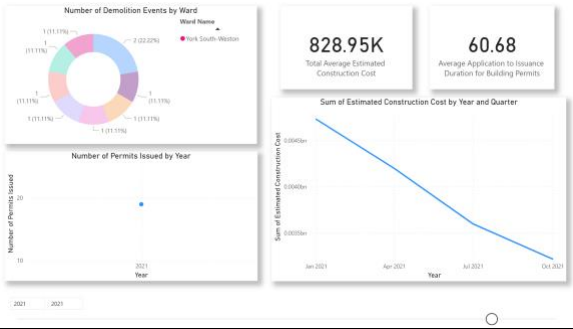
Dashboard page before slicing and dicing:



After dicing for the ward Beaches-East York and years between 2015-2021



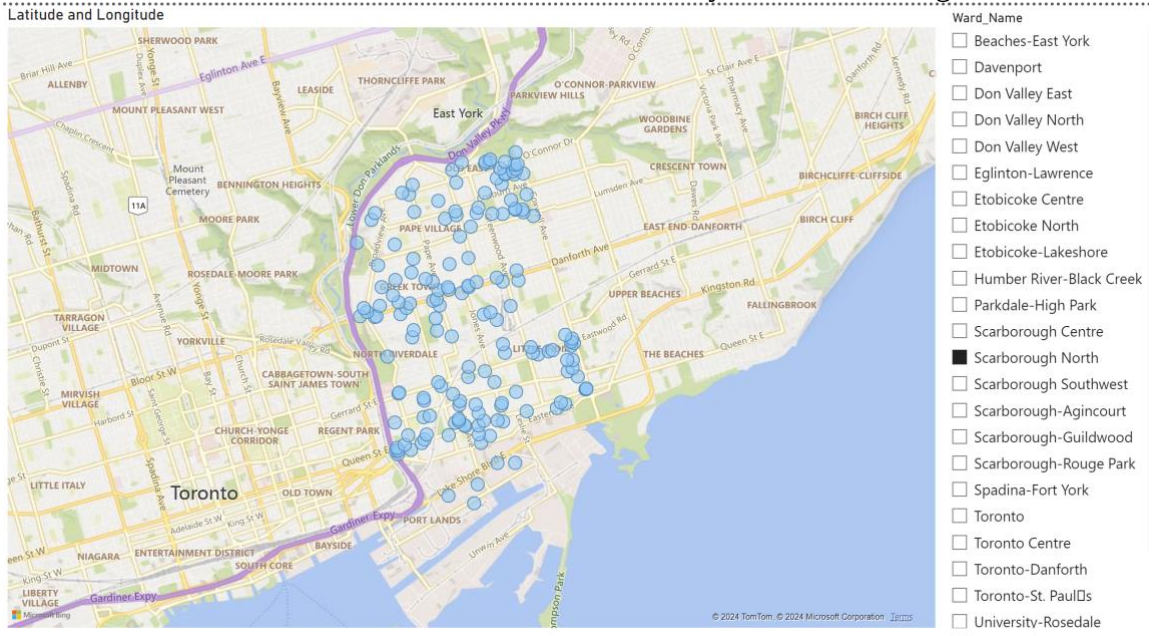
After dicing for the ward Etobicoke North and year 2021:



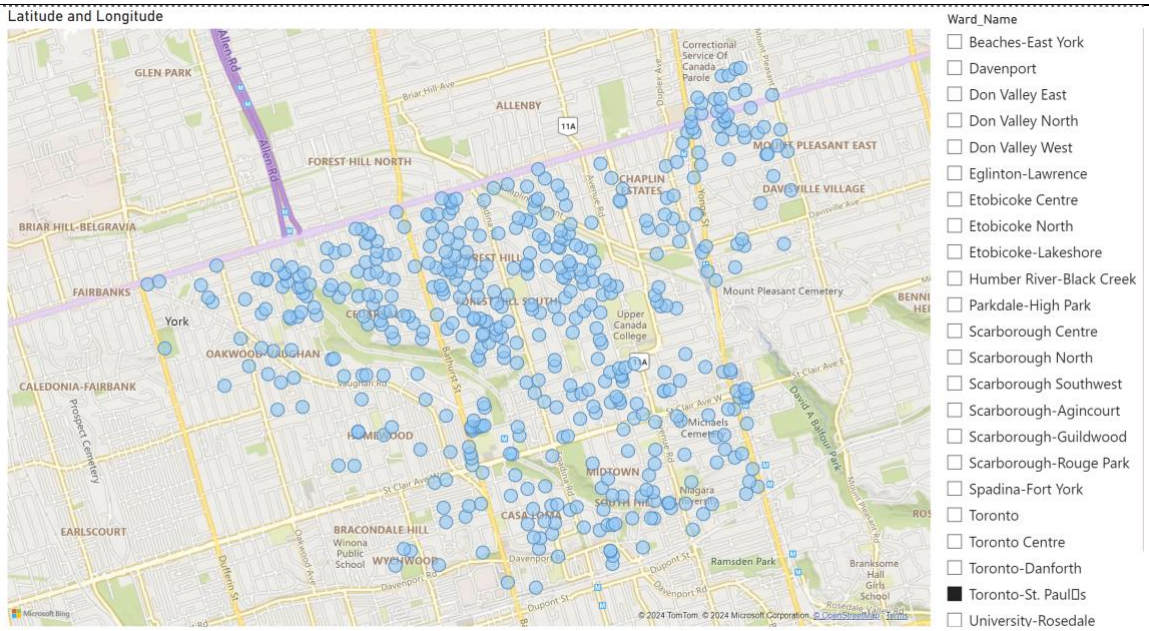
Executing Top N and Bottom N Queries

To visualize areas of increased construction activity, we plotted coordinates to map visualizations in Power BI. Below, all 8000+ construction projects are plotted, showing a trend of increased construction activity in central Toronto.

These projects can be divided by Toronto wards to explore trends in construction at the ward level. In the screenshot below, the coordinates are sliced by ward Scarborough North:



Here, they are sliced by Toronto St. Paul's:



Using the Power BI map visualization, we geocoded the coordinates to plot points on the map. The team created longitude and latitude columns by splitting the coordinates column, which were used in their respective ‘longitude’ and ‘latitude’ fields in the Power BI map visualization.

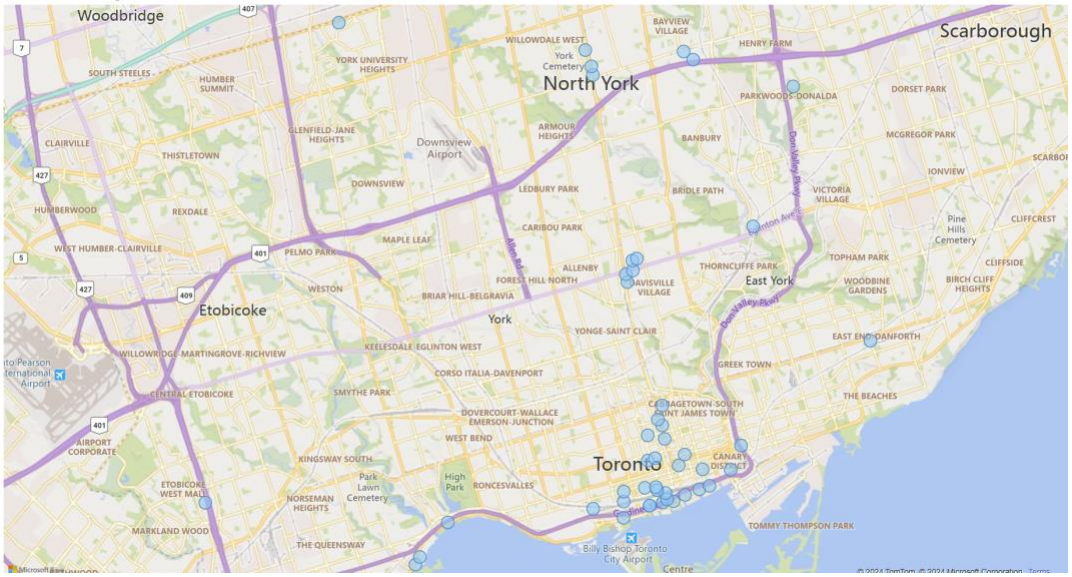
Power BI didn't have the Top N filter option available for map type visualizations, so we decided to create a calculated table which only included the coordinates of the top 50 estimated construction costs. This was achieved with a DAX expression:

```
Top 50 Building Permits Coordinates Table =
VAR Top50Costs =
    TOPN(
        50,
        ALL('development_fact_table'),
        'development_fact_table'[Est_Const_Cost],
        DESC
    )
VAR Top50PermitKeys = SELECTCOLUMNS(
    Top50Costs,
    "Permit_Key", 'development_fact_table'[Permit_Key]
)
VAR Top50Details =
    SELECTCOLUMNS(
        FILTER(
            ALL('BuildingPermitsDimension'),
            'BuildingPermitsDimension'[Permit_Key] IN Top50PermitKeys
        ),
        "Permit_Key", 'BuildingPermitsDimension'[Permit_Key],
        "Longitude", 'BuildingPermitsDimension'[Longitude],
        "Latitude", 'BuildingPermitsDimension'[Latitude],
        "Ward_ID", RELATED('development_fact_table'[Ward_ID])
```

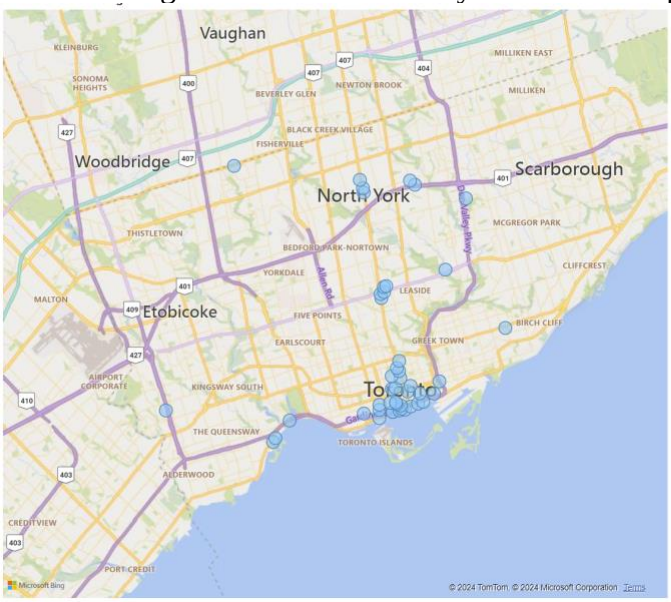
```
    )
RETURN
    Top50Details

```

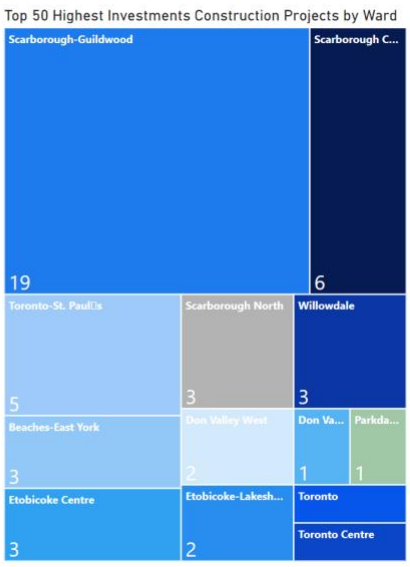
A relationship between the Ward Dimension table and the Top 50 Building Permit Coordinates table was created, with the linking column being Ward ID. This was to ensure that visuals were able to be filtered by Ward Name for user understandability and clarity.

The results of the top 50 query were a much cleaner map visual, allowing the user to get a clearer picture of where the highest investment construction activity is.

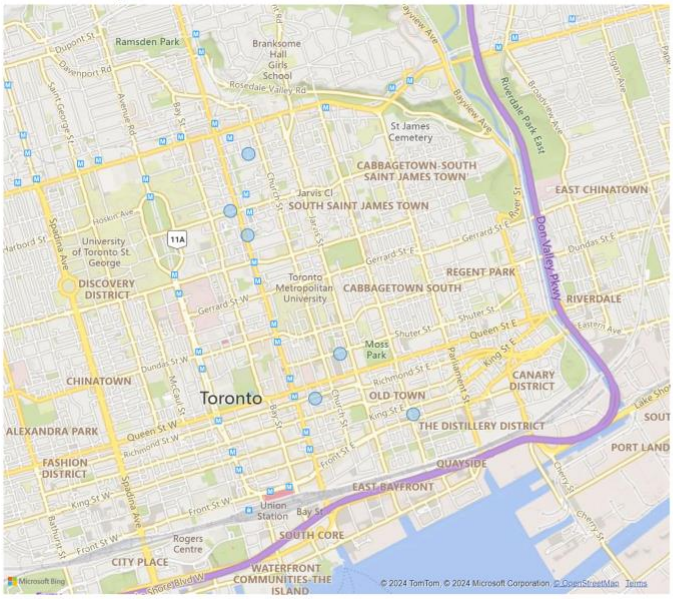


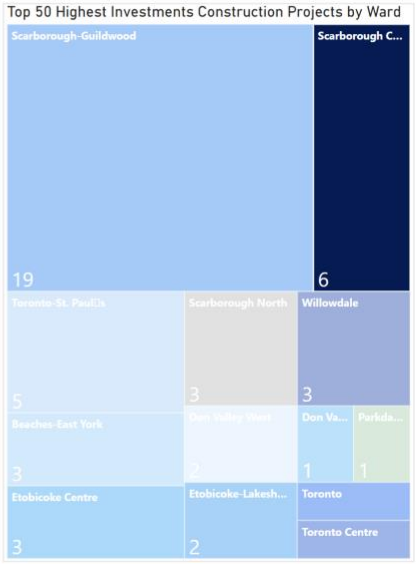
To provide a clearer picture of the breakdown of the top 50 highest construction cost projects, the team included a tree map showing the count of permits organized by each ward on the dashboard. This gives the user the ability to slice the map visualization by ward(s) seamlessly.



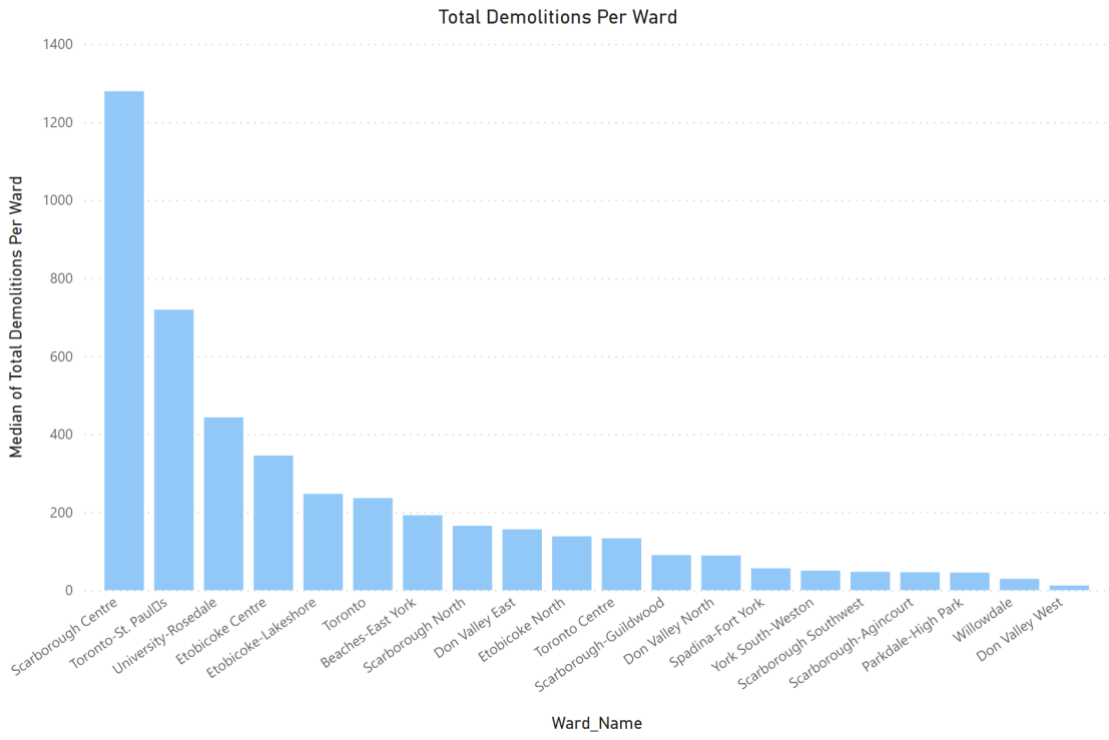


By clicking on one of the wards in the tree map, for example, Scarborough Center, the user can view the high-investment projects mapped to only this ward. In the below screenshot, only the 6 construction projects of Scarborough Center can be seen on the map.

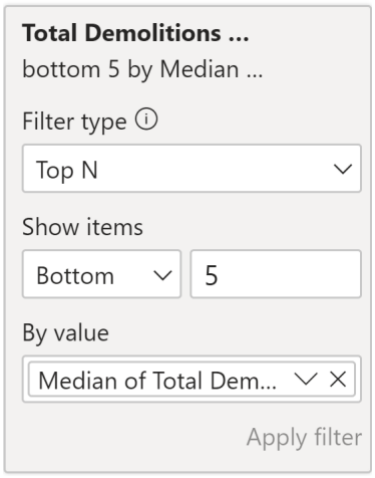




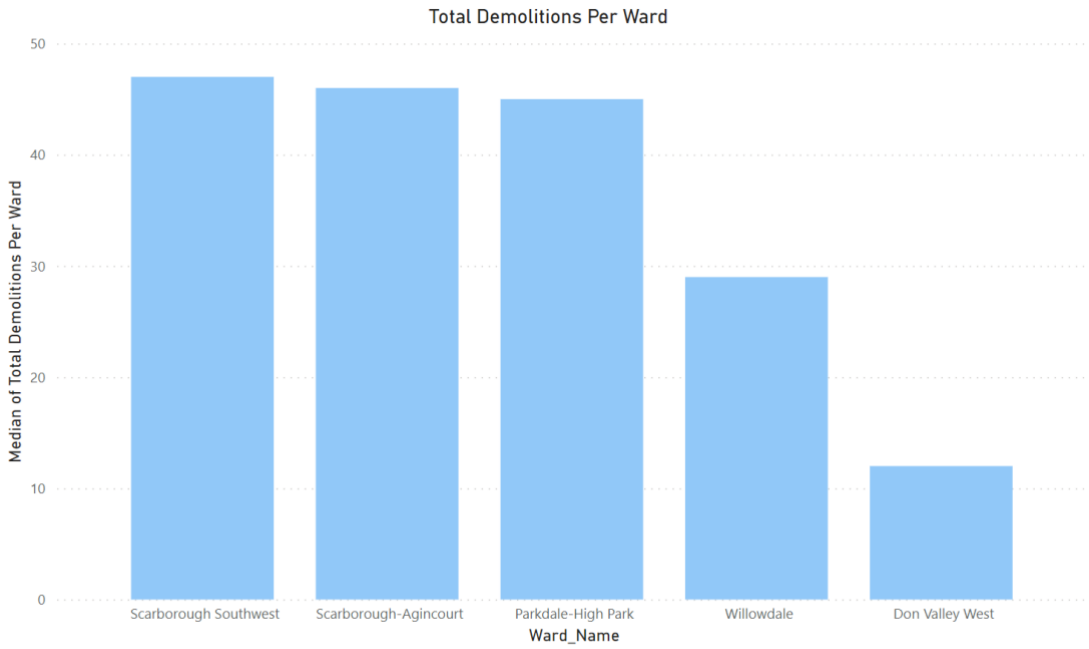
Using the "total demolition" fact in the demolition fact table, a visual displaying the total demolitions per ward was created.



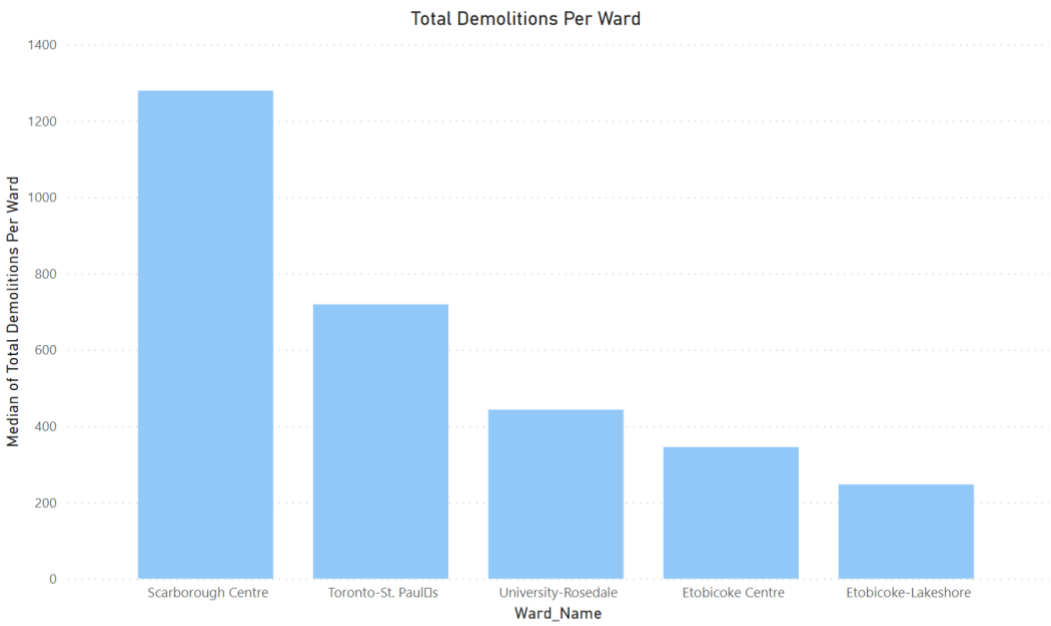
To identify the wards with the lowest demolition activity, we can use a bottom N query to find the bottom 5 wards in terms of demolition rates. This was achieved using Power BIs Top N filter type.



The resulting visualization filtered for the bottom 5 wards is shown below.



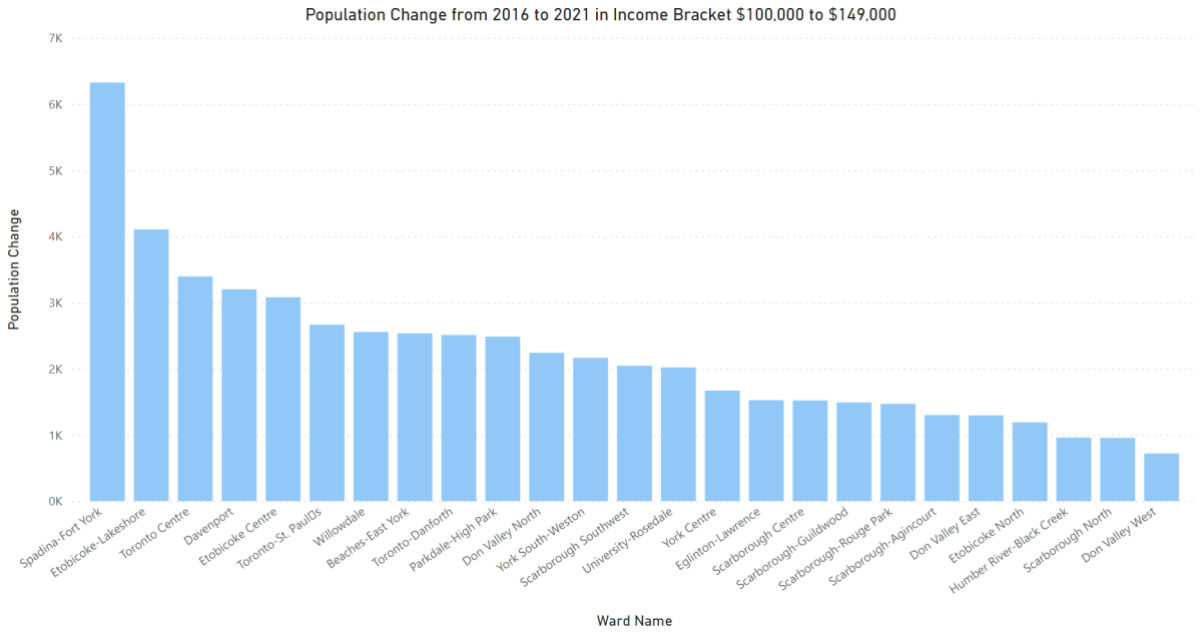
The visual was also filtered to display the Top 5 wards with the highest demolition rates, as seen in the screenshot below. This is particularly relevant in the context of our analysis, as high rates of demolition can be a sign of gentrification making way for new developments.



After observing the results from the top N query for wards with the highest demolition rates, as well as the top N query for wards with the highest construction investment activity, the intersection of wards for both query results is the Scarborough Center, Etobicoke Center, and Etobicoke Lakeshore.

Our team wanted to explore the change in the population in terms of income. Specifically, we wanted to examine the change in population from 2016 to 2021 in terms of the $100,000 to $149,000 income bracket to see in which wards the greatest increase in wealth was taking place.

The visualization below shows the population increase for each ward for the $100,000 to $149,000 income bracket.



This was achieved with the following DAX expression:

```
Population_Change =
CALCULATE (
    SUM (ward_profile_fact_table[Population]),
    ward_profile_fact_table[Year] = 2021,
    IncomeDimension[Income] = "$100,000 to $149,999"
) - CALCULATE (
    SUM (ward_profile_fact_table[Population]),
    ward_profile_fact_table[Year] = 2016,
    IncomeDimension[Income] = "$100,000 to $149,999"
)

```

To filter the above visualization for the top 5 wards, we created a DAX expression to rank the wards by the increase in population:

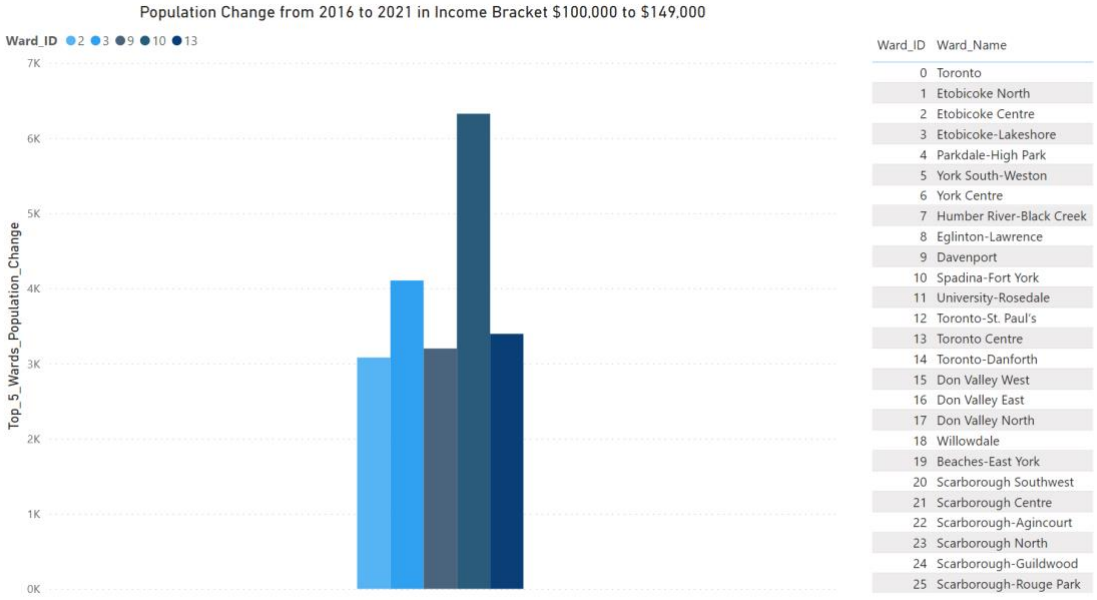
```
Rank_Wards_by_Population_Change =
RANKX (
    ALL (ward_profile_fact_table[Ward_ID]),
    [Population_Change],
    ,DESC,
    Dense
)

```

Our Top N query is shown below to filter to get the top 5 ranks.

```
Top_5_Wards_Population_Change =
IF (
    [Rank_Wards_by_Population_Change] <= 6,
    [Population_Change],
    BLANK()
)
```

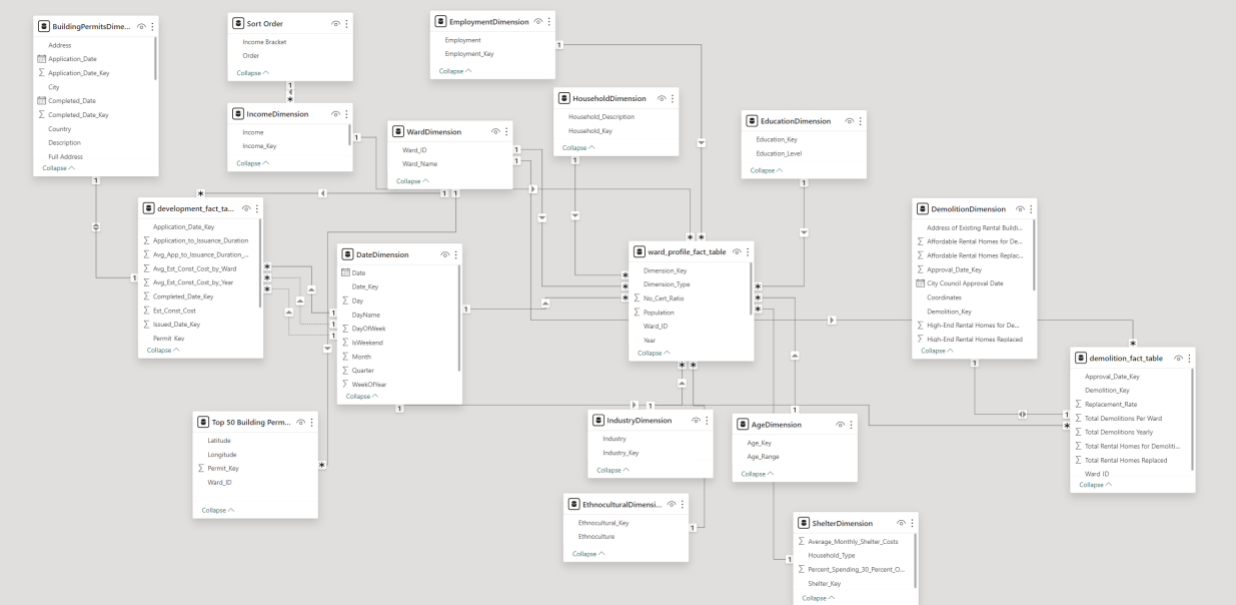
*Note that "<=6" is due to Toronto having a Ward_ID of 0, which we excluded from the analysis to focus on the wards themselves.

Our visualization below displays the top 5 wards with the highest increase in the population in income bracket $100,000 to 149,000.



Ward_ID	Ward_Name
0	Toronto
1	Etobicoke North
2	Etobicoke Centre
3	Etobicoke-Lakeshore
4	Parkdale-High Park
5	York South-Weston
6	York Centre
7	Humber River-Black Creek
8	Eglinton-Lawrence
9	Davenport
10	Spadina-Fort York
11	University-Rosedale
12	Toronto-St. Paul's
13	Toronto Centre
14	Toronto-Danforth
15	Don Valley West
16	Don Valley East
17	Don Valley North
18	Willowdale
19	Beaches-East York
20	Scarborough Southwest
21	Scarborough Centre
22	Scarborough-Agincourt
23	Scarborough North
24	Scarborough-Guildwood
25	Scarborough-Rouge Park

Power BI Datamart Design



*Note that Sort Order table was created to sort the x-axis for income brackets visualization in a custom order