Northeastern University

DAMG 7290 Spring 2022 - TEAM 8

Analysis of Boston Airbnb Data

Group Members

➢ Adarsh Chidanand Havalad

➢ Deep Kothari

➢ Karthik Donthineni

➢ Monil Rawka

Analysis of Boston Airbnb Data

Table of Contents:

1. Introduction :

Airbnb has established itself as a new leader in the hospitality business. The main concept is to discover a way for locals to earn some extra cash by renting out space to visitors to the area.

The company offers services through its app and website, which collects a large amount of data about the hosts and visitors. We'll use the information to look at the number of listings, the number of reviews for the listings, the descriptions of the homes, availability, average occupancy, and more.

The data is collected from insideairbnb.com website and has the overall data divided into 3 categories such as Listings, Reviews and Calendar. The other dataset contains statistics on the population and median income of certain zip codes.Using these two datasets we will try to draw insights into how Airbnb's business is affected by the location as well as understand the effect Airbnb is having on the neighborhood itself. The csv files are loaded and transformations will be made in the staging area and visual analysis will be done using business intelligence tools.

2. Dataset Description:

- The 'listings' dataset is the primary dataset which has 74 attributes and 3349 records. Some of the important attributes in the dataset are:

Field	**Type**	**Description**
id	BIGINT	Airbnb's unique identifier for the listing
listing_url	nvarchar(MAX)	URL of the listing
host_id	BIGINT	Airbnb's unique identifier for the host/user
host_since	date	The date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest
host_is_superhost	nvarchar(MAX)	t=true, f=false
host_listings_count	BIGINT	The number of listings the host has
neighborhood_cleansed	nvarchar(MAX)	Neighborhood in which the

		listing is located
property_type	nvarchar(MAX)	Self selected property type. Hotels and Bed and Breakfasts are described as such by their hosts in this field
room_type	nvarchar(MAX)	[Entire home/apt|Private room|Shared room|Hotel]
bathrooms_text	nvarchar(MAX)	The number of bathrooms in the listing
bedrooms	BIGINT	The number of bedrooms
price_in_dollars	BIGINT	Daily price in local currency
has_availability	nvarchar(MAX)	t=true,f=false
availability_x	BIGINT	The availability of the listing x days in the future.
number_of_reviews	Float	The number of reviews the listing has
review_scores_parameter	Float	The average rating of the listing for a parameter like cleanliness, location etc.
reviews_per_month	Float	The number of reviews the listing has over the lifetime of the listing

- The second dataset is the ‘reviews’ which has 6 attributes and 128,311 records.

Field	**Type**	**Description**
listing_id	BIGINT	Airbnb ID of the listing
date	date	Date the review was published
reviewer_id	BIGINT	Airbnb’s unique identifier for the host/user
comments	nvarchar(MAX)	The content of the review

- The third dataset is the ‘calendar’ which has 7 attributes and 1,048,576 records.

Field	Type	Description
listing_id	BIGINT	Airbnb ID of the listing
date	date	The date of the listing
price_in_dollars	BIGINT	Daily price in local currency
minimum_nights	BIGINT	minimum number of night stay for the listing (calendar rules may be different)
maximum_nights	BIGINT	maximum number of night stay for the listing (calendar rules may be different)

- The final dataset is the ‘income by zip codes’ which is 3 attributes and 520 records.

Field	Type	Description
Median Household Income in Dollars	BIGINT	Median household income in local currency
ZIP	BIGINT	ZIP code
Population	BIGINT	Population of the ZIPcode

3. Data Source Mapping:

The following table shows how some of the attributes from the source system are mapped to the attributes in the target system.

- Fact Table

Source Table and Attribute	**Target Table and Attribute**
listing.id	dbo.Facttable.id
listing.host_id	dbo.Facttable.host_id
zipcode.zip	dbo.Facttable.zipcode

- Host Details

Source Table and Attribute	**Target Table and Attribute**
listing.host_url	dbo.Dim_Host_Details.host_url
listing.host_name	dbo.Dim_Host_Details.host_name
listing.host_city	dbo.Dim_Host_Details.host_city
listing.host_country	dbo.Dim_Host_Details.host_Country

Availability Details

Source Table and Attribute	**Target Table and Attribute**
listing.has_availability	dbo.Dim_Availability_Details.has_availability
listing.availability_30	dbo.Dim_Availability_Details.availability_30
listing.availability_60	dbo.Dim_Availability_Details.availability_90
listing.availability_90	dbo.Dim_Availability_Details.availability_60
listing.availability_365	dbo.Dim_Availability_Details.availability_365

- Property Details

Source Table and Attribute	Target Table and Attribute
listing.neighborhood_cleansed	dbo.Dim_Property_Details.neighbourhood_cleansed
listing.property_type	dbo.Dim_Property_Details.property_type
listing.bedrooms	dbo.Dim_Property_Details.bedrooms
listings.bathrooms_text	dbo.Dim_Property_Details.bathrooms_text

- Review Details

Source Table and Attribute	Target Table and Attribute
reviews.id	dbo.Dim_Reviews_Details.Review_id
reviews.date	dbo.Dim_Reviews_Details.date
reviews.reviewer_name	dbo.Dim_Reviews_Details.reviewer_name
reviews.comments	dbo.Dim_Reviews_Details.Comments

- Income Details

Source Table and Attribute	Target Table and Attribute
zip code.rank	dbo.Dim_Income_Details.Rank
zipcode.Zipcode	dbo.Dim_Income_Details.Zipcode
zip code.population	dbo.Dim_Income_Details.Population
zipcode.median_income_in_dollars	dbo.Dim_Income_Details.median_income

- Review Ratings Details

Source Table and Attribute	Target Table and Attribute
listing.first_review	dbo.Dim_Reviews_Rating_Details.First_review
listing.review_score_rating	dbo.Dim_Reviews_Rating_Details.review_score_rating
listing.reviews _per_month	dbo.Dim_Reviews_Rating_Details.reviews.per_month

4. Entities and Attributes:

All the attributes of Dimension Tables and Fact tables are as follows

1) Dim_Property_Details - property_id (Primary Key) id, Bedrooms, listing_url, neighbourhood_cleansed, property_type, accommodates, bathrooms_text, bedrooms, price

2) Dim_Reviews_Ratings_Details - Rating_id (Primary Key), id, First_review, Last_review,review_score_rating,review_score_accuracy, review_score_cleanliness,review_score_value,review_score_value, reviews_per_month

3) Dim_Availability_Details - Availibility_id(PrimaryKey) , has_availability,availability_30, has_availability_60, has_availability_90, has_availability_365

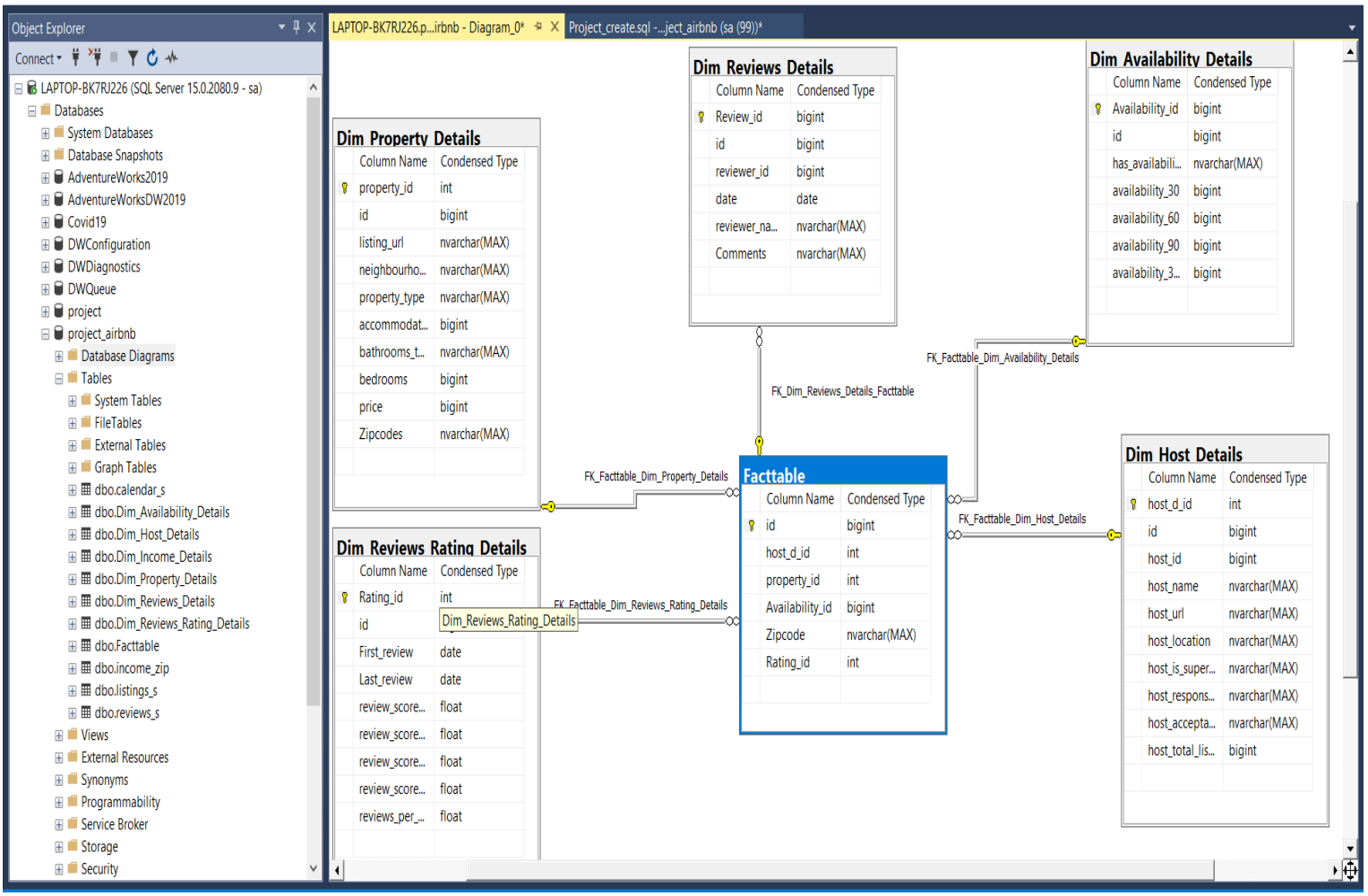
4) Dim_Host_Details - host_d_id (Primary Key), host_id, host_name, host_url, hosT_location,host_is_superhost,host_response_rate, host_acceptance_rate, host_total_listings_count

5) Dim_Income_Details - Zipcode (Primary Key), Rank, Population, median_income

6) Dim_Reviews_Details - Review_id (Primary Key), id, reviewer_id, date, reviewer_name, Comments

7) Fact_Table - id, host_id, property_id, Availability_id, Zipcode, Rating_id

5. Database Diagram:



6. Implementation

Before loading the data into a database management software which in our case is SQL Server, we first start by processing the flat files to make the process of loading as smooth as possible without any errors. Since we are loading a csv file we check to see if any of the values in the dataset have commas or any other special characters in them which could disrupt our load and cause errors such as truncation errors. We replace those characters with other characters such as an underscore or dash or whatever is deemed appropriate for that field. We then use python to create a zip code column using the latitude and longitude columns and assign them to the listings.

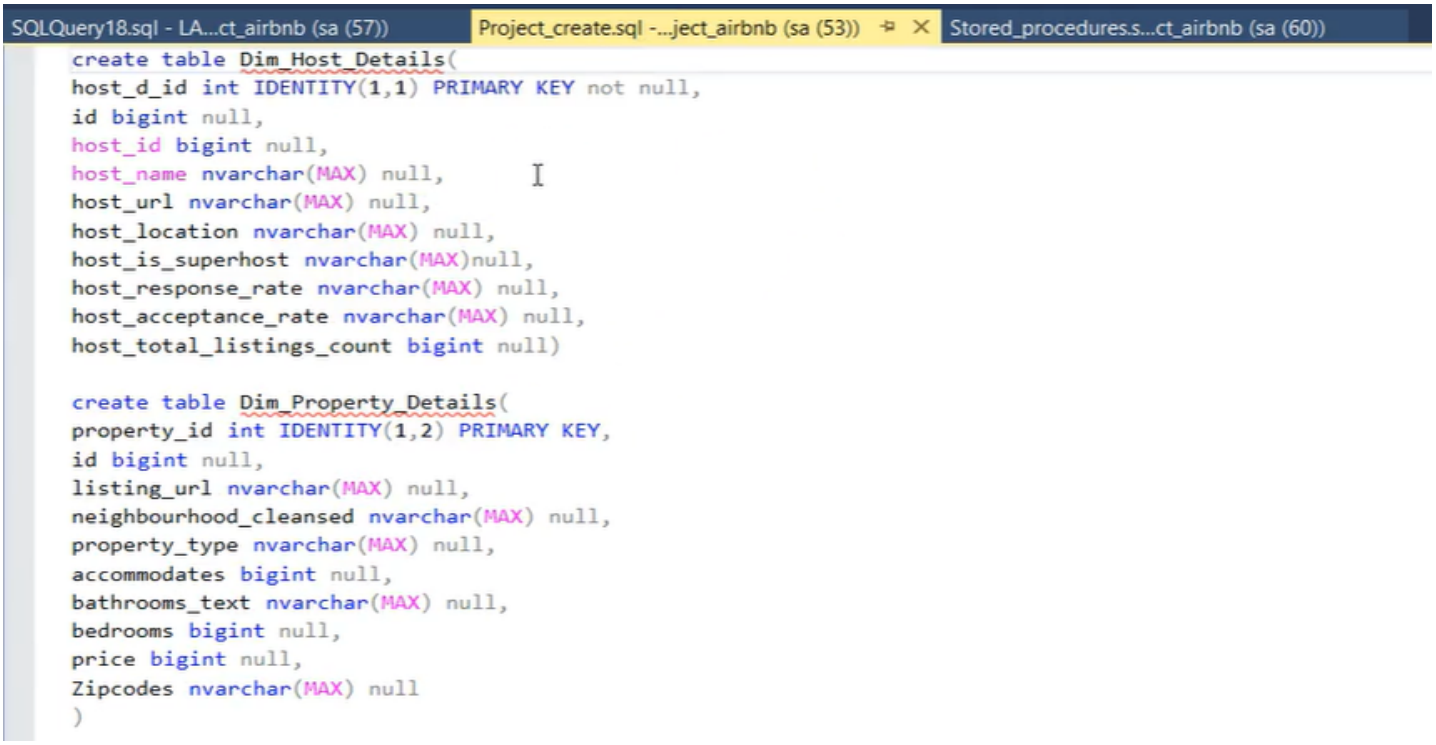
6.1 Initial load of the staging data

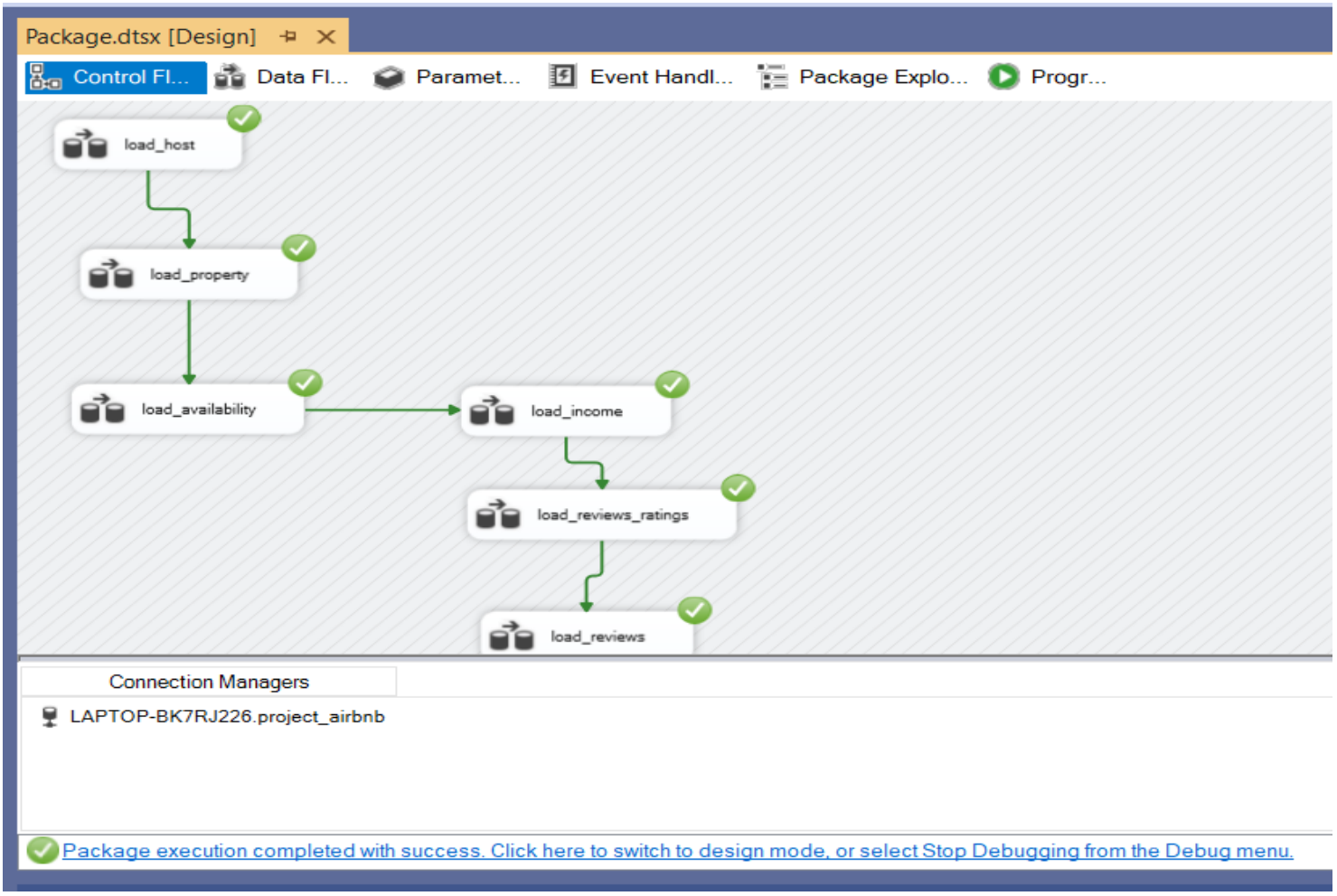
We first use the flat-file import export wizard to load our csv files - listings, reviews and calendar, zip code. These will be our staging tables using which we will build out our warehouse.

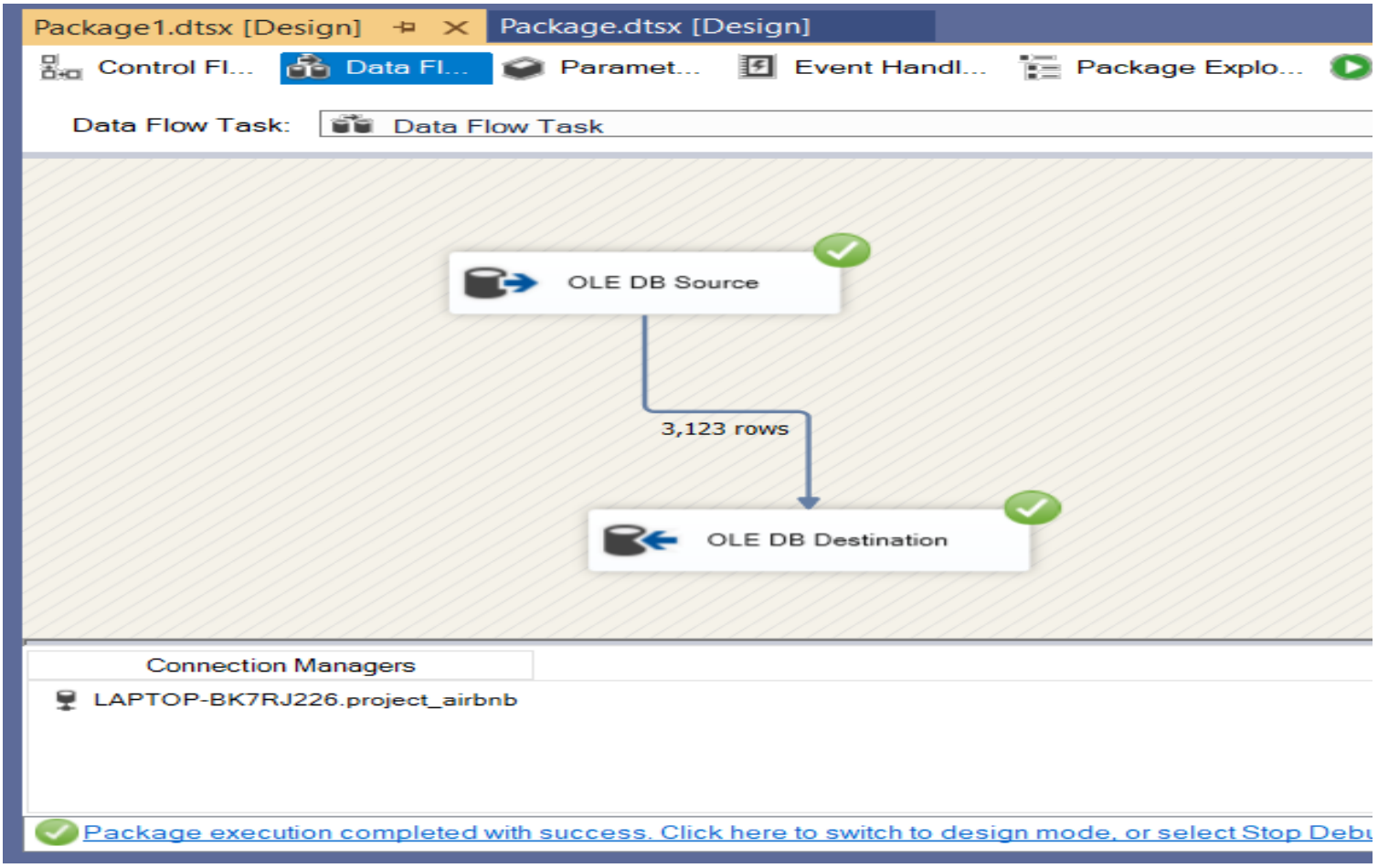
6.2 Initial loading of the dimension tables

We first create the dimension tables we need by referring to our schema and then writing the necessary create statements in SQL server management studio. After we create the tables, we use the create diagram function in SQL server to connect the tables and implement primary and foreign keys.

Once the empty tables are created we use an SSIS package to populate them with the source being our staging table which we imported earlier. We have multiple data flows each corresponding to a particular dimension table and inside the data flows we simply take the necessary columns from the staging table and map them to our destination tables.

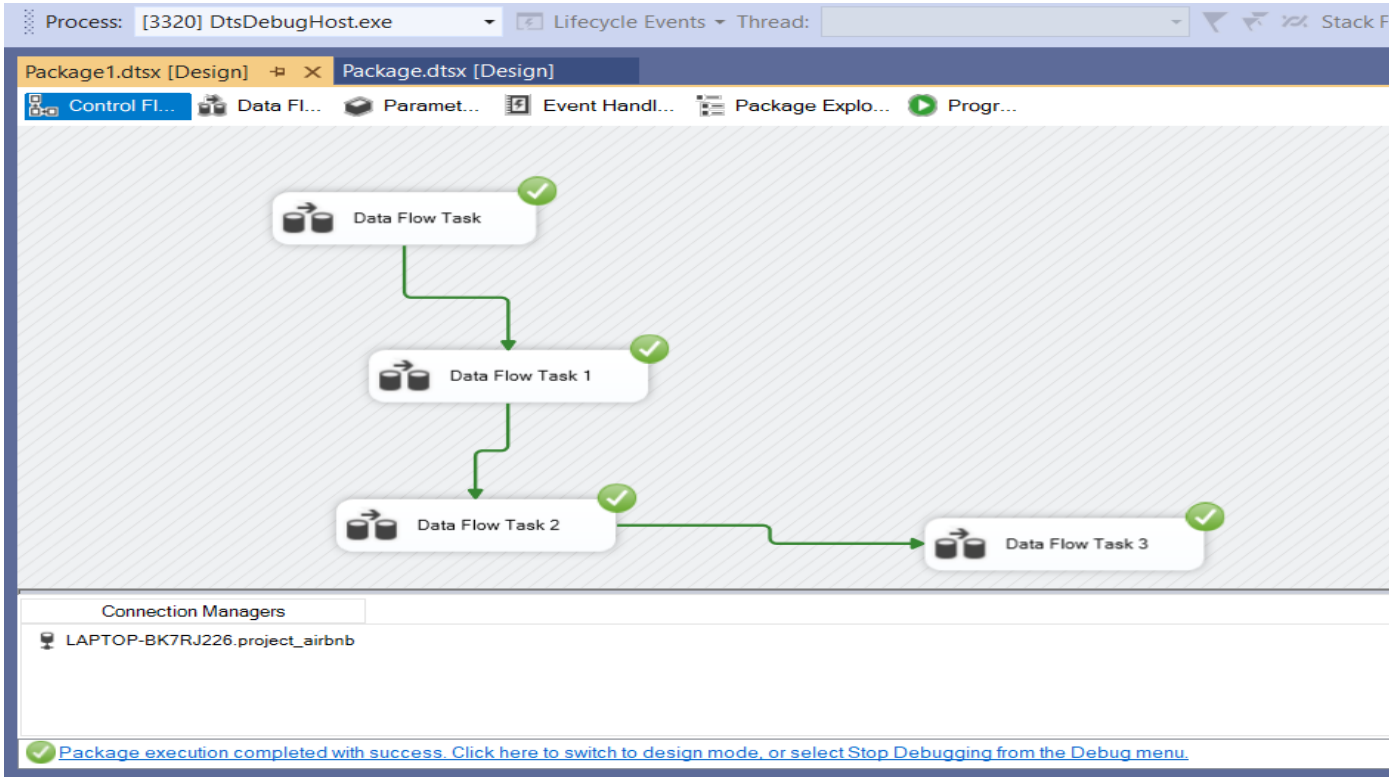


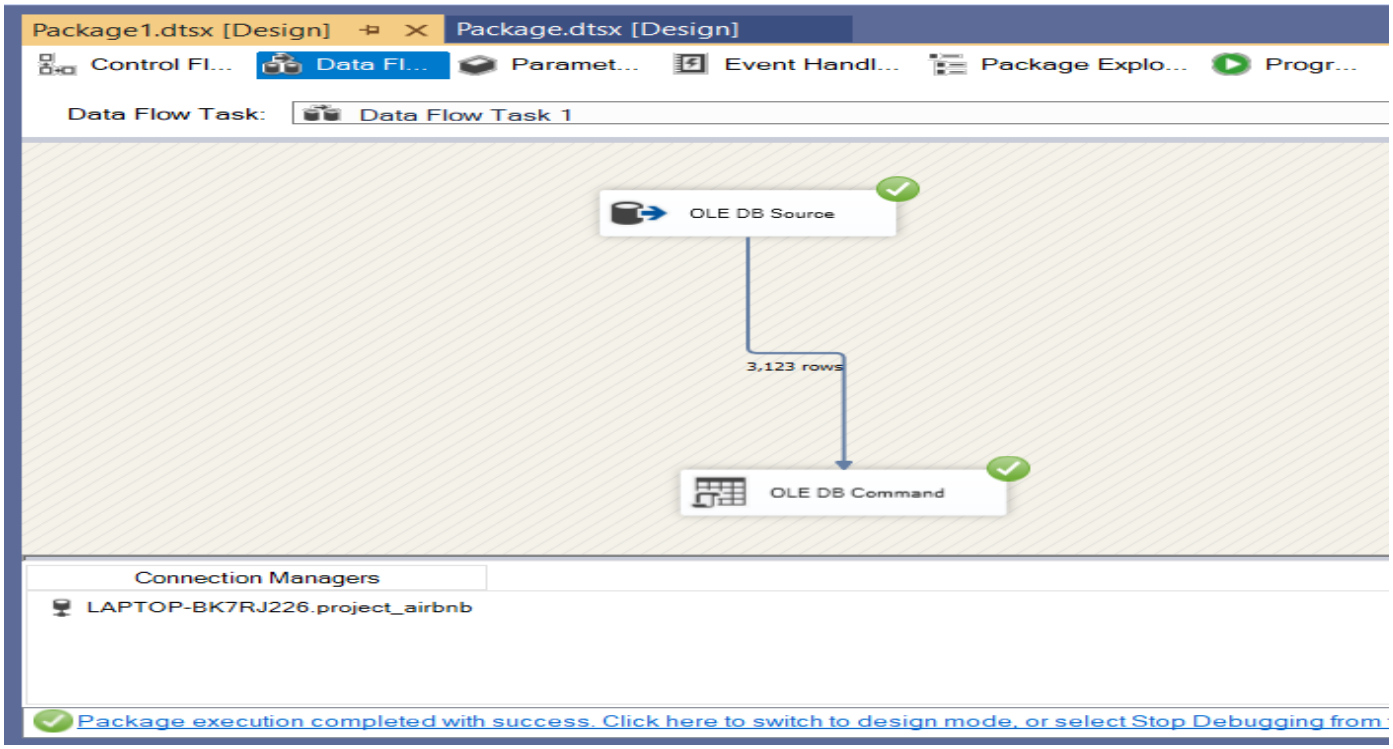

Package.dtsx [Design]
Control Fl...
Data Fl...
Paramet...
Event Handl...
Package Explo...
Progr...
load_host
load_property
load_availability
load_income
load_reviews_ratings
load_reviews
Connection Managers
LAPTOP-BK7RJ226.project_airbnb
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.


Package1.dtsx [Design]
Package.dtsx [Design]
Control Fl...
Data Fl...
Paramet...
Event Handl...
Package Explo...
Data Flow Task:
Data Flow Task
OLE DB Source
3,123 rows
OLE DB Destination
Connection Managers
LAPTOP-BK7RJ226.project_airbnb
Package execution completed with success. Click here to switch to design mode, or select Stop Debu

6.3 Initial loading of the fact table

After the dimension tables are created we use them to populate the fact table with the necessary key columns. We use the same control and data flow structure as above.



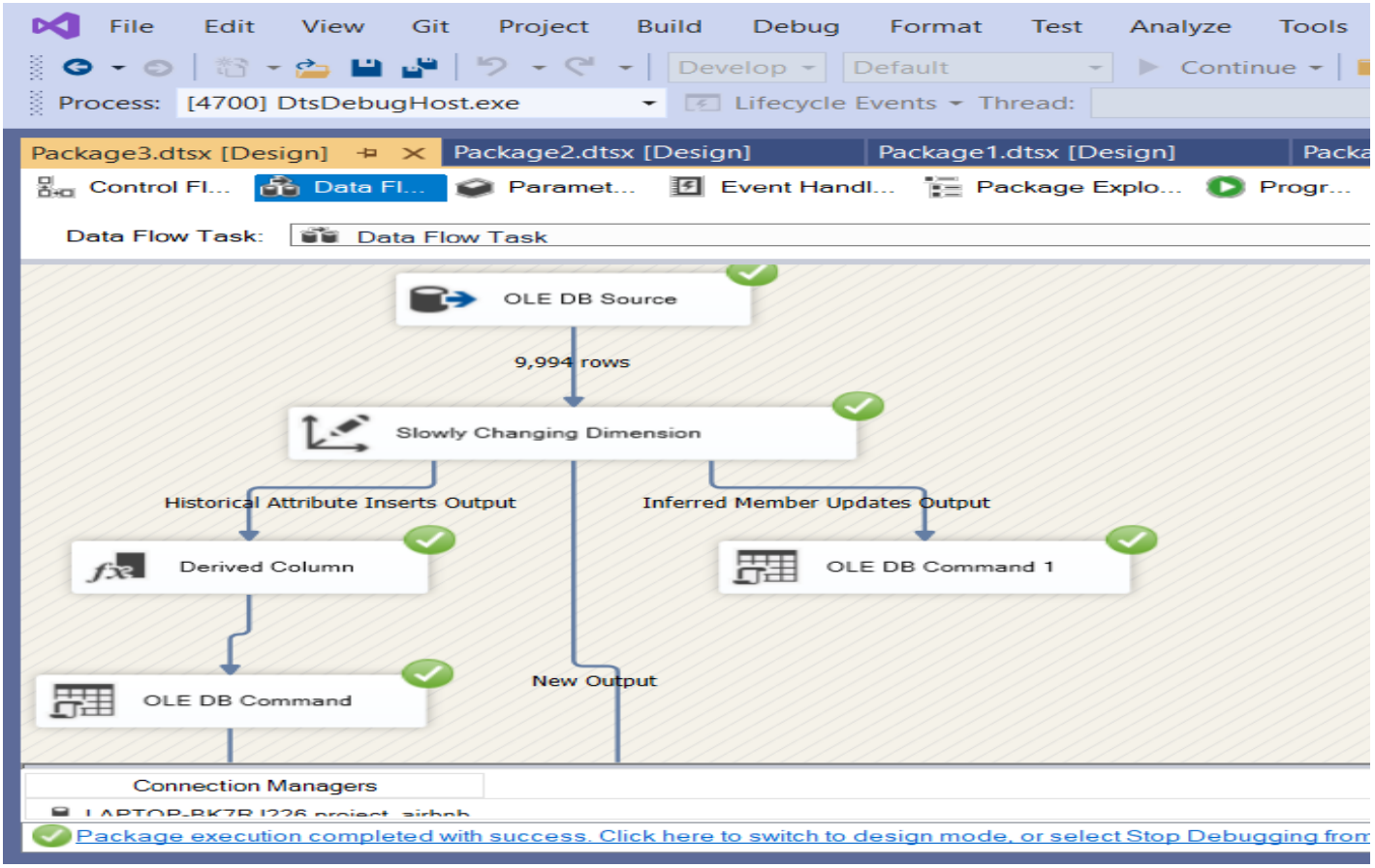


6.4 Incremental loading of next quarter data

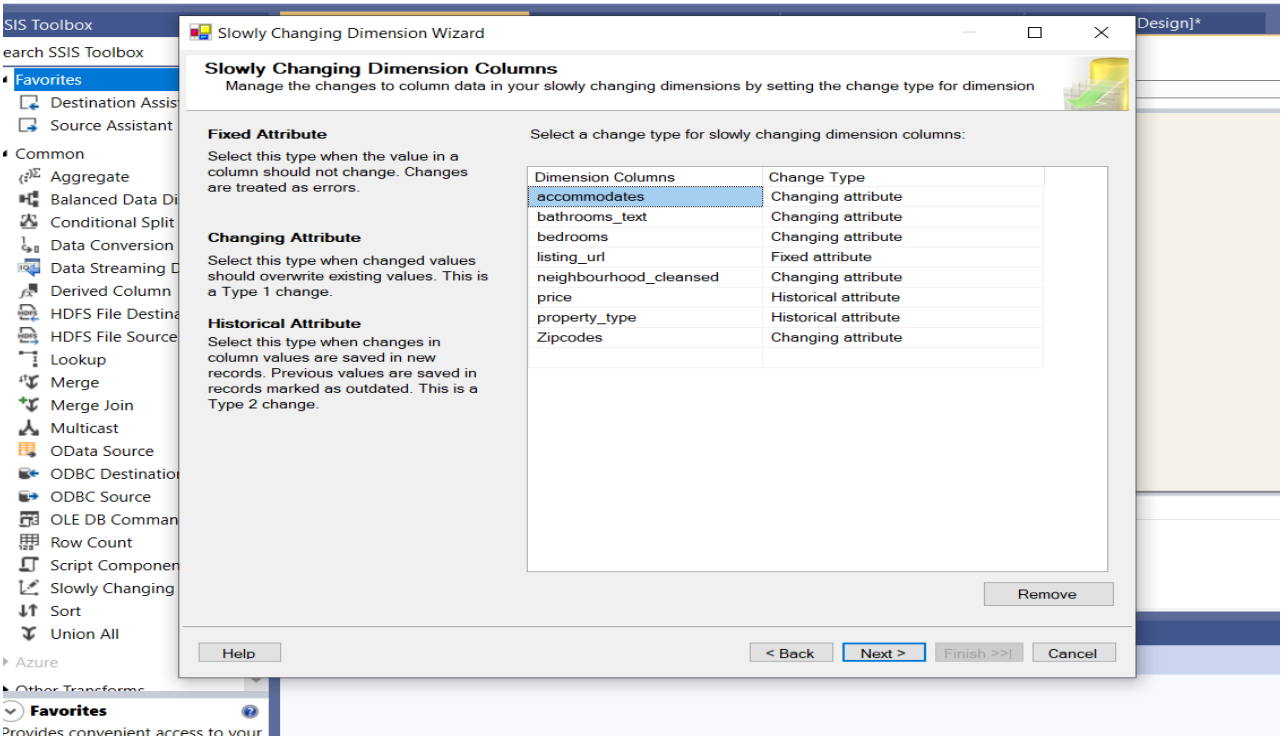
We use the import wizard again to import the december data file and make the december staging table. We do not update the september staging table but instead create a new one because in our case instead of just getting new records, the entire csv with the new records and the updated records is published every quarter. We keep the September staging table for preserving history.

6.4.1 Assignment of SCD types to Columns in Dimension Tables

We then modified the dimension tables to assign them SCD types that we need.



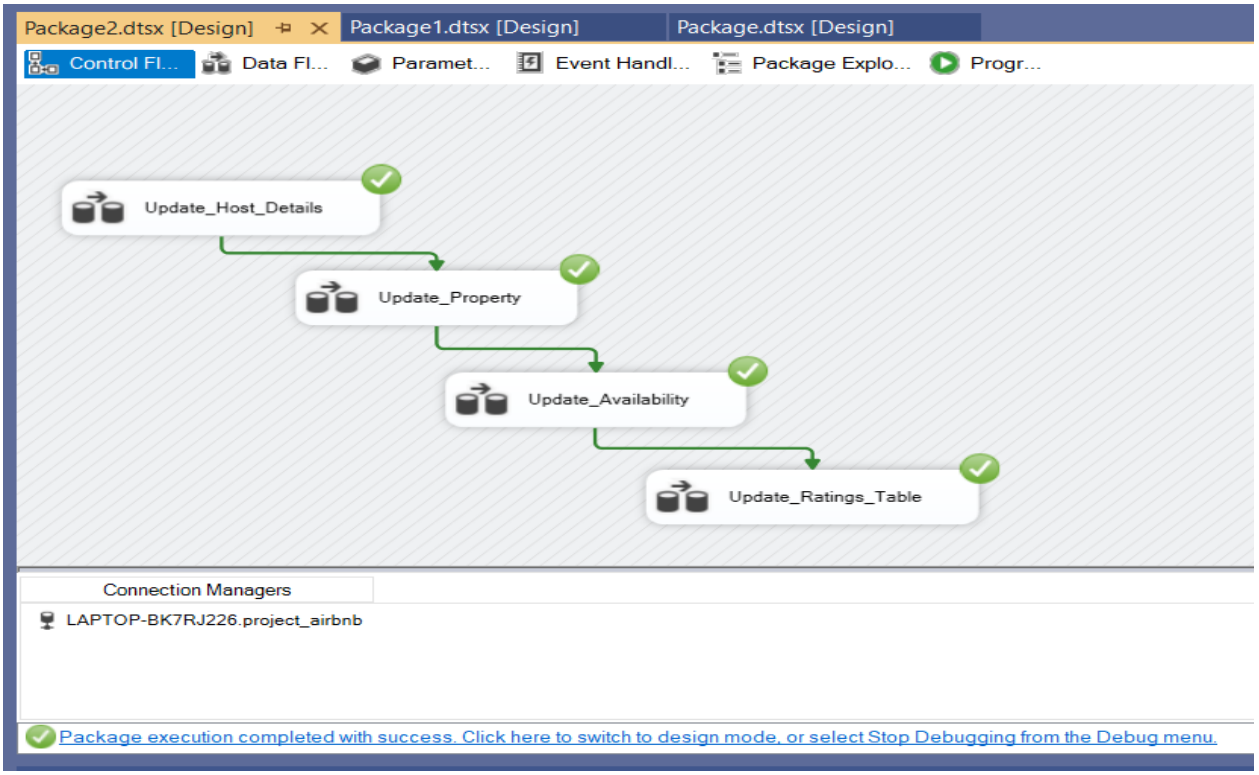
Assigning SCD

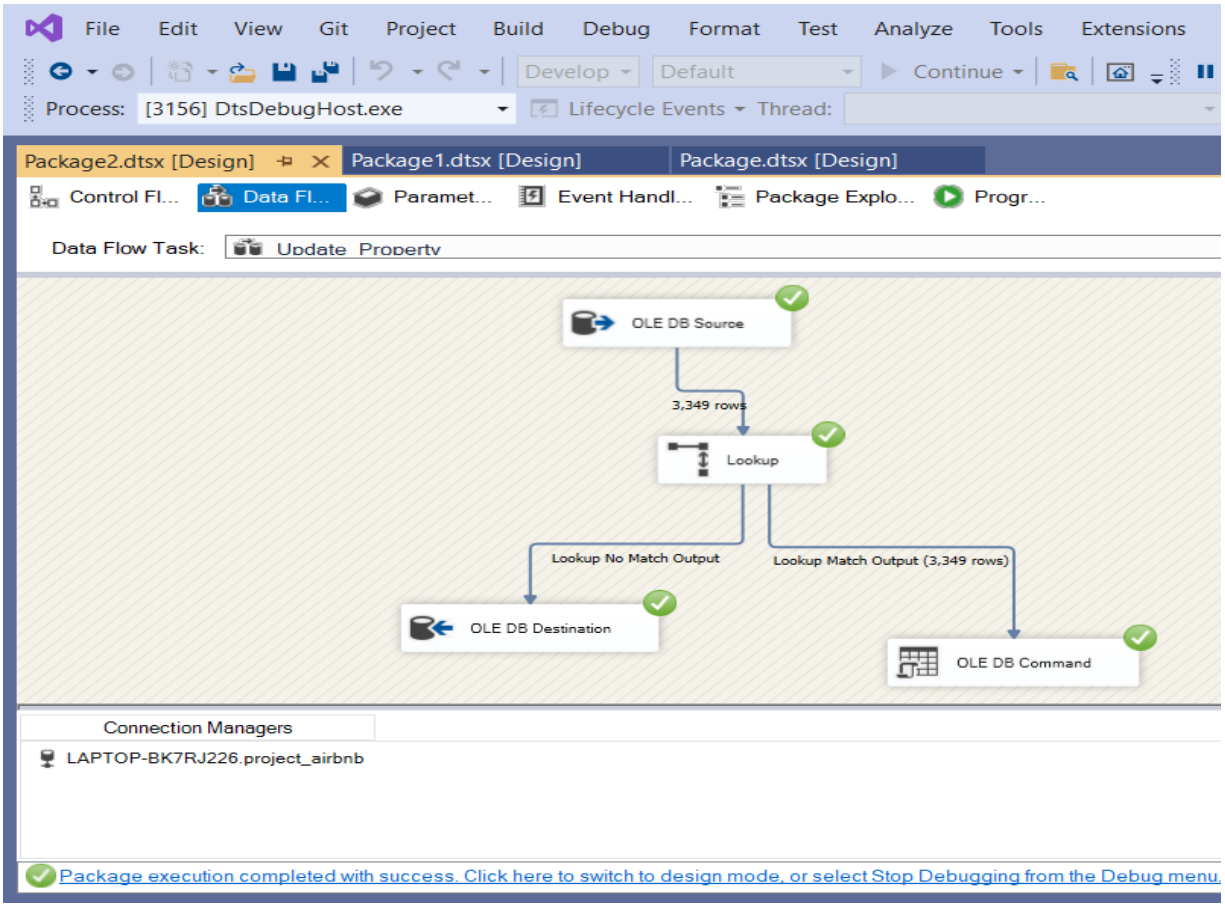


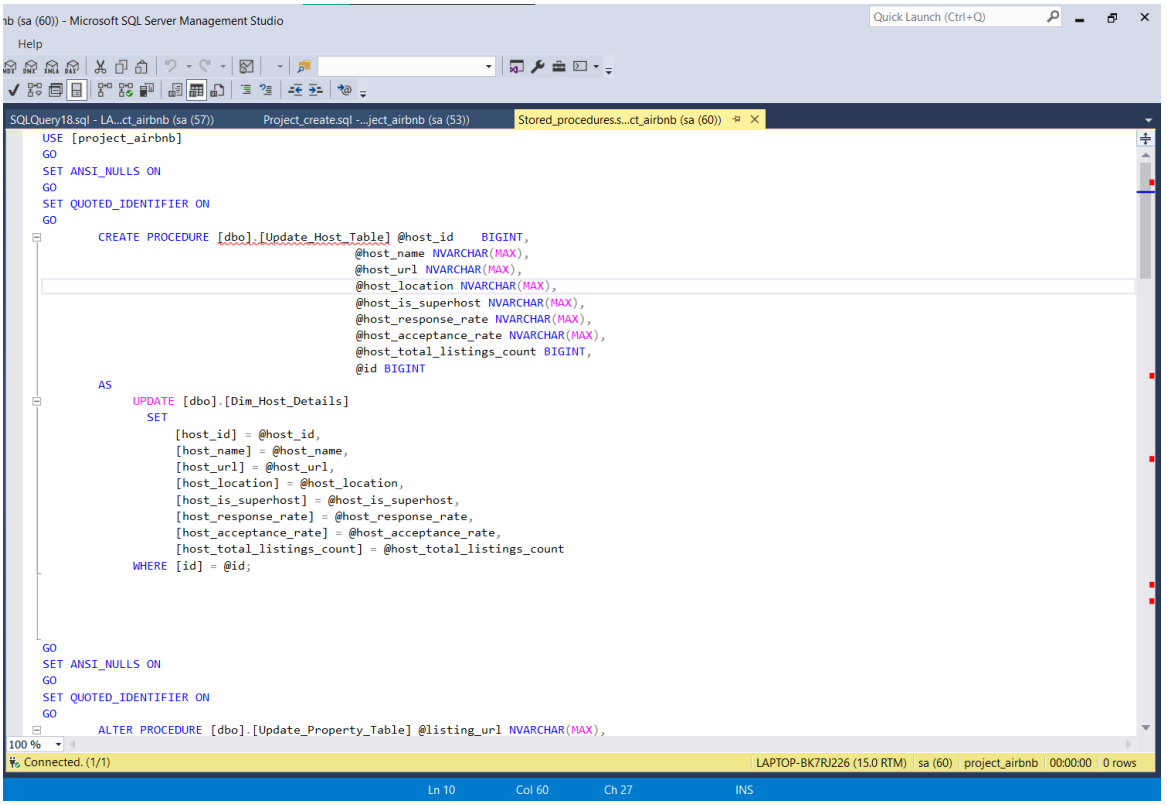
Assigning types

6.4.2 Updating the dimension and fact tables with the new data

After the December staging table is loaded in we use it to insert new records and update old ones in our dimension table. We create a new SSIS package and in the control flow we insert multiple data flows and connect them.Within each data flow we take the december staging table as source and do a lookup with the destination dimension table, the resulting no match output is used to insert new records to the dimension while the match output is used to update the records based on SCD types. To use the match output to update records we need to use the OLE DB command and feed it the executed stored procedure statement which would be written in the database.


Package2.dtsx [Design]
Package1.dtsx [Design]
Package.dtsx [Design]
Control Fl...
Data Fl...
Paramet...
Event Handl...
Package Explo...
Progr...
Update_Host_Details
Update_Property
Update_Availability
Update_Ratings_Table
Connection Managers
LAPTOP-BK7RJ226.project_airbnb
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.


File
Edit
View
Git
Project
Build
Debug
Format
Test
Analyze
Tools
Extensions
Develop
Default
Continue
Process:
[3156] DtsDebugHost.exe
Lifecycle Events
Thread:
Package2.dtsx [Design]
Package1.dtsx [Design]
Package.dtsx [Design]
Control Fl...
Data Fl...
Paramet...
Event Handl...
Package Explo...
Progr...
Data Flow Task:
Update_Property
OLE DB Source
3,349 rows
Lookup
Lookup No Match Output
Lookup Match Output (3,349 rows)
OLE DB Destination
OLE DB Command
Connection Managers
LAPTOP-BK7RJ226.project_airbnb
Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.



These are the procedures used to update the data.

7. Business Intelligence

Business intelligence (BI) is the process of transforming data into actionable insights that aid businesses and individuals in making strategic decisions and comprehending business. We used tableau, a business intelligence application, to derive insights from our data in our project. We've set up a connection between the SQL server and Tableau and created a few data visualizations.

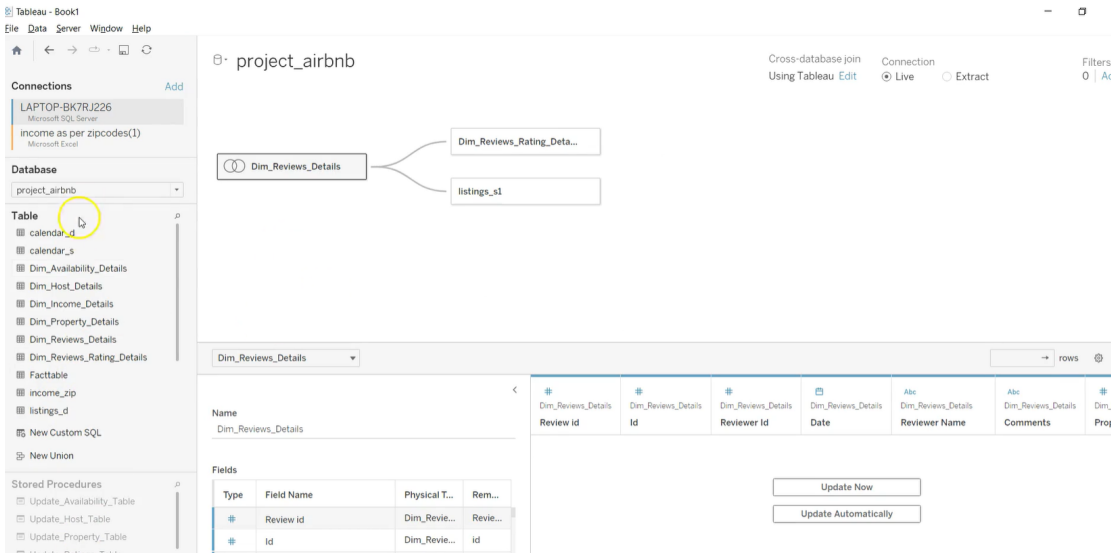
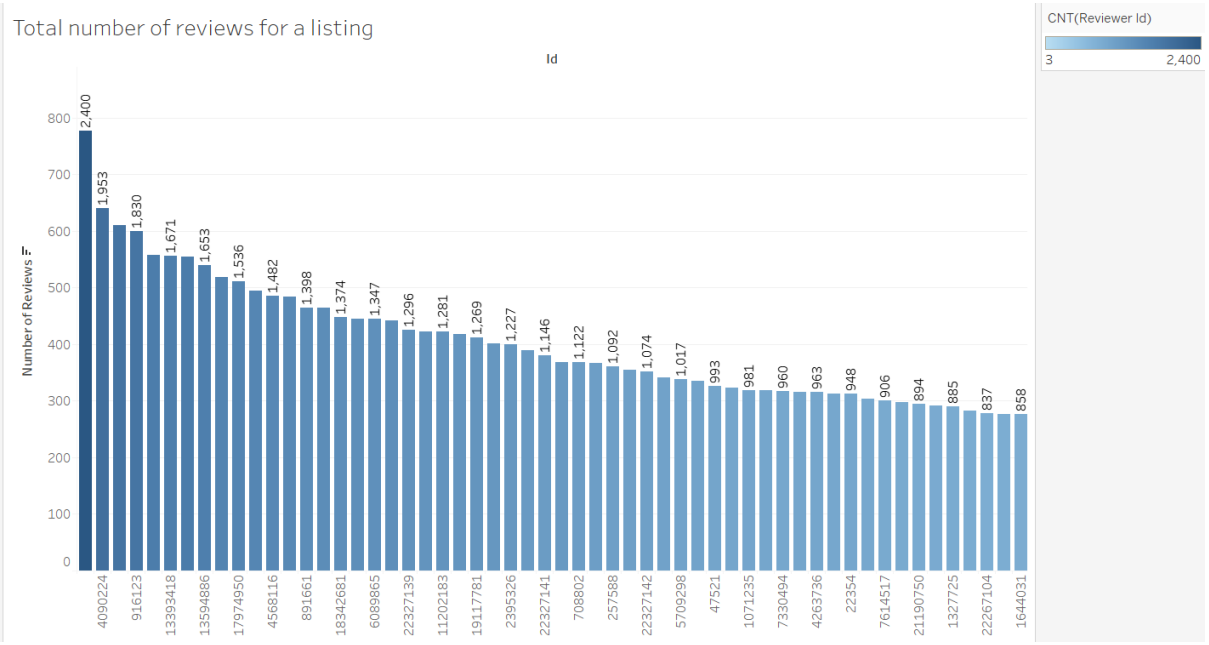


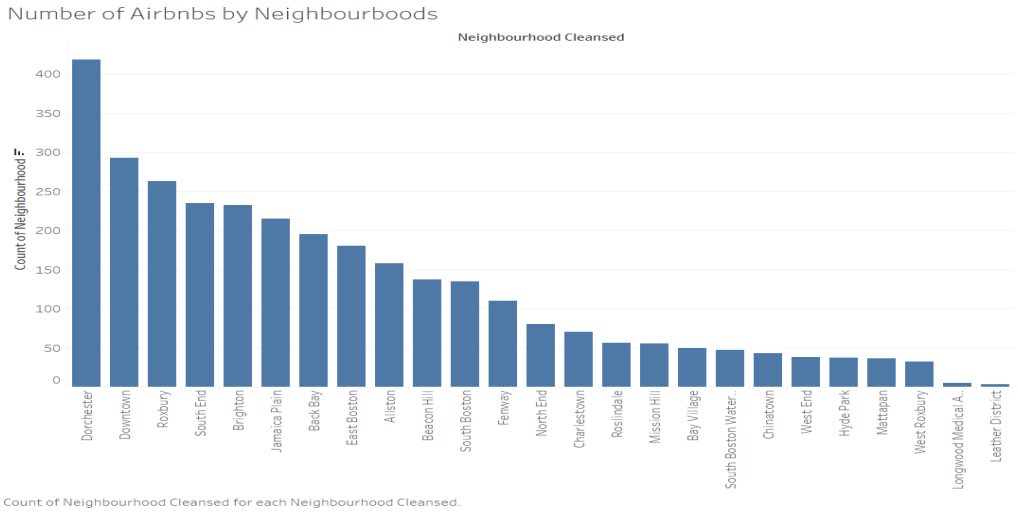
Tableau connection to the database

1) Number of reviews for the listings



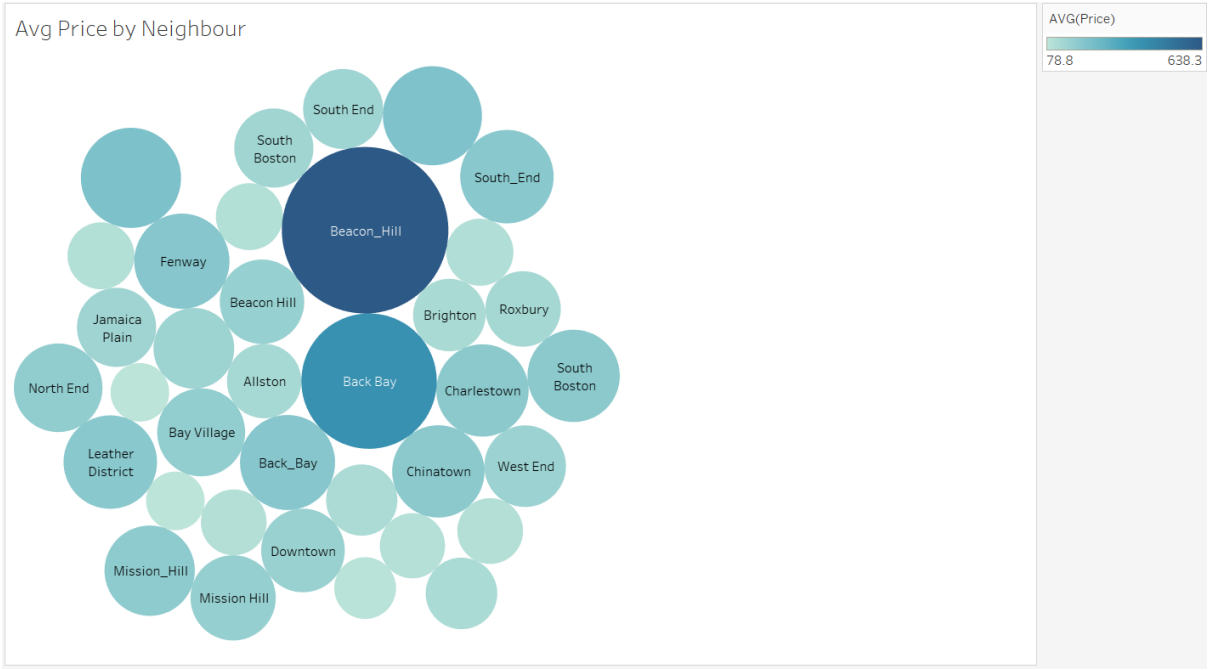
Here we can observe that the listings id with the highest count of 2400 reviews is leading and this can help in shortlisting the property by looking through the reviews.

2) Number of Airbnbs by Neighborhoods in Boston



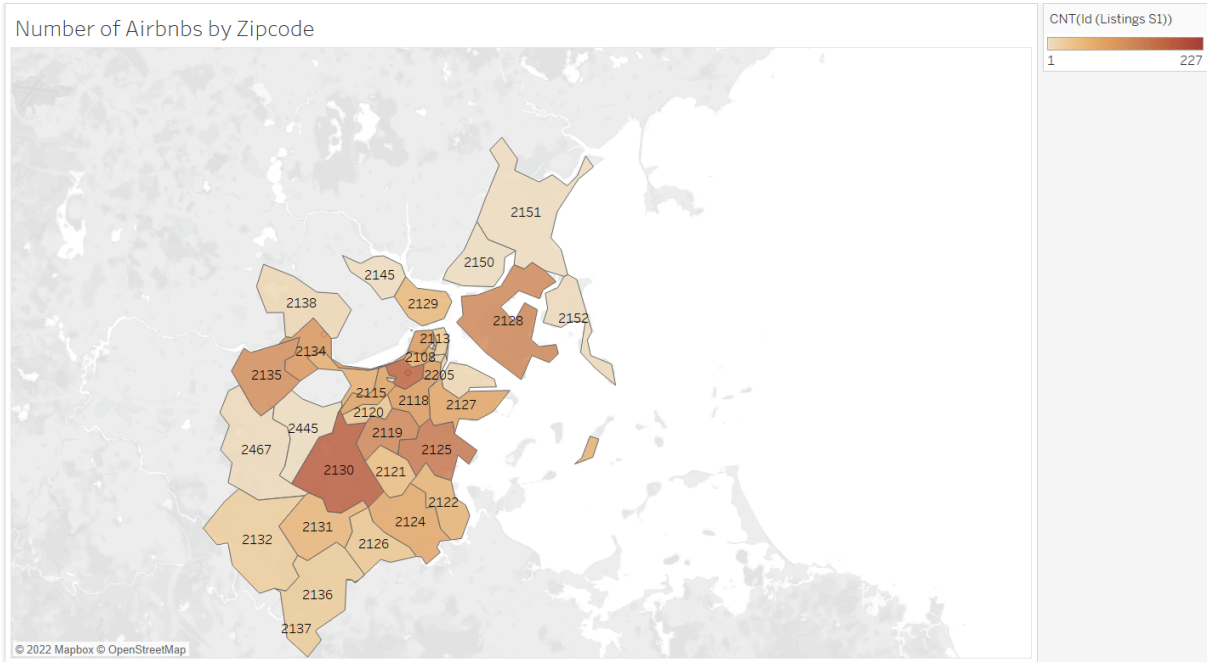
As we can observe, the maximum number of Airbnbs are in the Dorchester neighborhood of Boston, followed by Downtown and Roxbury.

3) Average Price of Airbnbs by Neighbourhoods in Boston



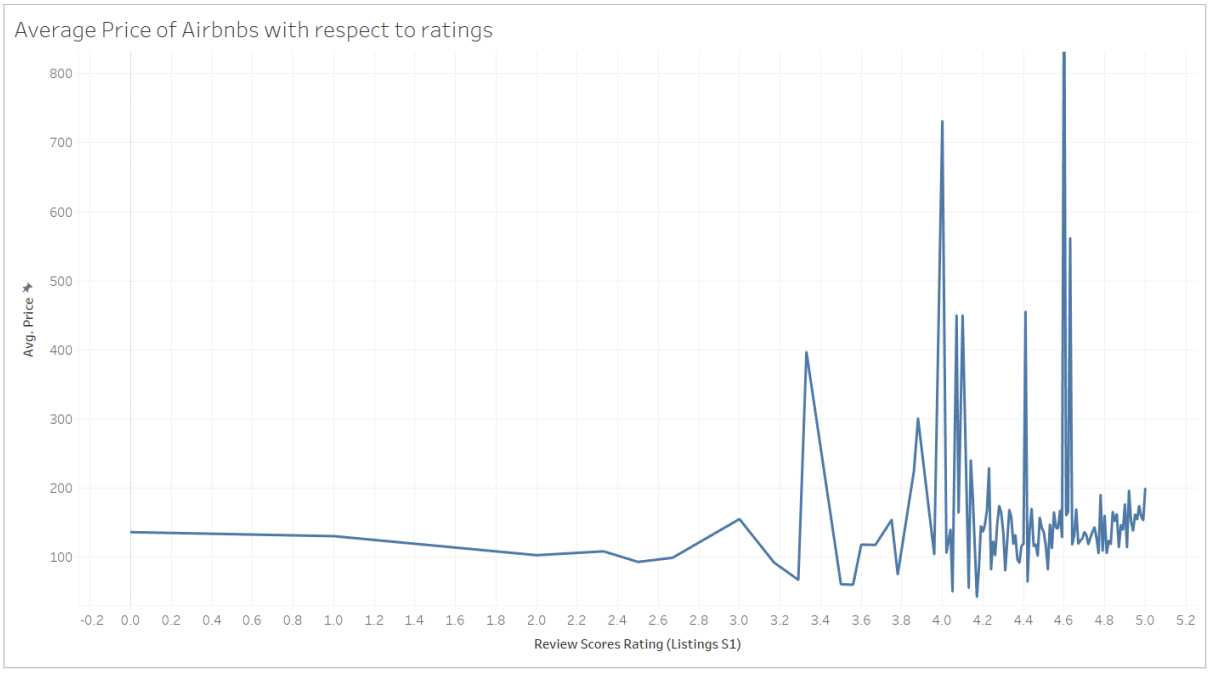
We can notice that the neighborhood ‘Beacon_Hill’ has the most expensive Airbnbs but as we can see from the bar chart in the previous visualization, it has very few listings, so that can be considered an outlier. BackBay can be considered the most expensive neighborhoods with a significant amount of listings.

4) Number of Airbnbs by ZIP Codes



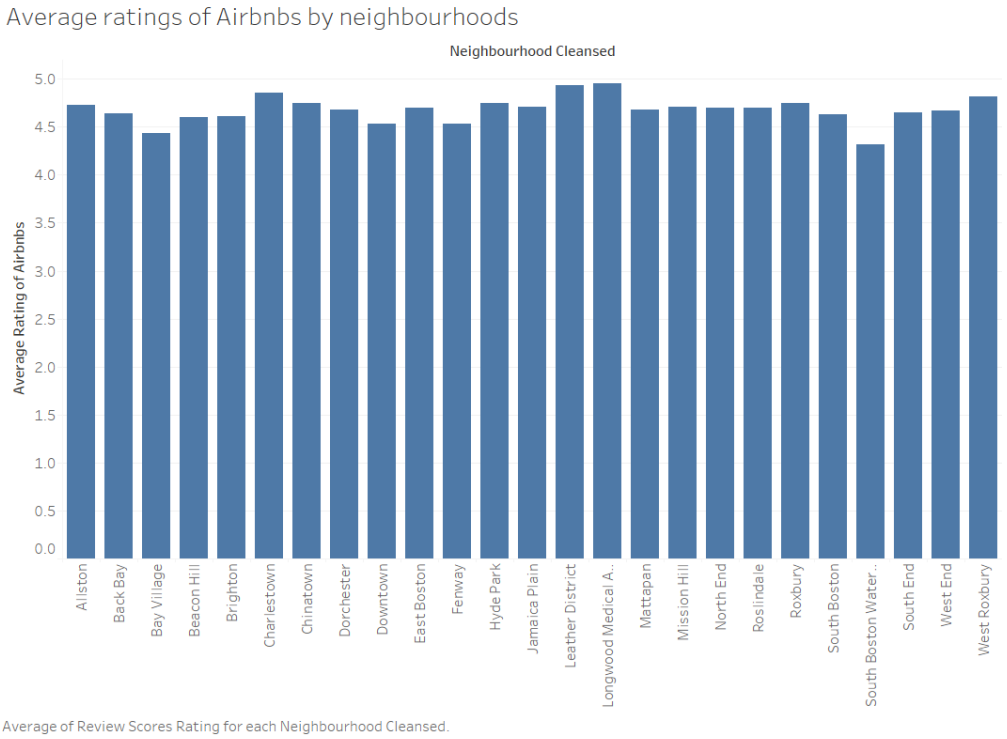
We see that most Airbnbs are located around central Boston and the zip code 02130 that is Jamaica Plain has the most number of airbnbs listed.

5) Average Price of Airbnbs with respect to Ratings



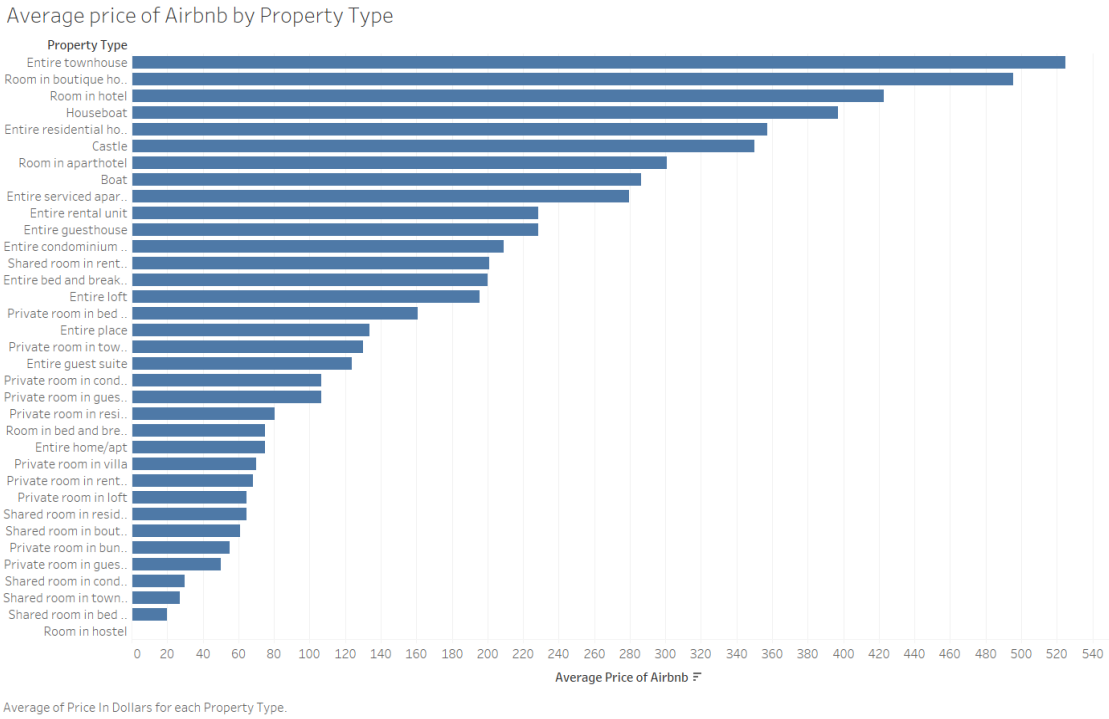
Apart from some outliers, we can see that as the average rating of a listing increases, the price of the same is higher and the average price for the boston airbnbs is about $160 -$200.

6) Average Ratings of Airbnbs by Neighborhoods in Boston



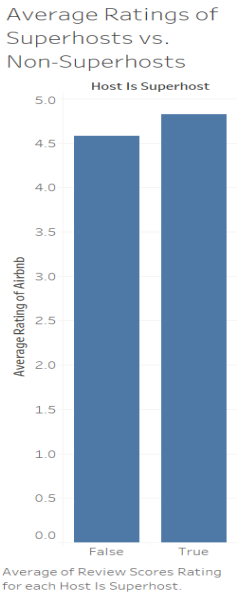
There does not seem to be a lot of variance in the average rating of the listings in different neighborhoods, which says that good properties can be found in Boston, regardless of the neighborhood.

7) Average Price of Airbnbs by Type of Property



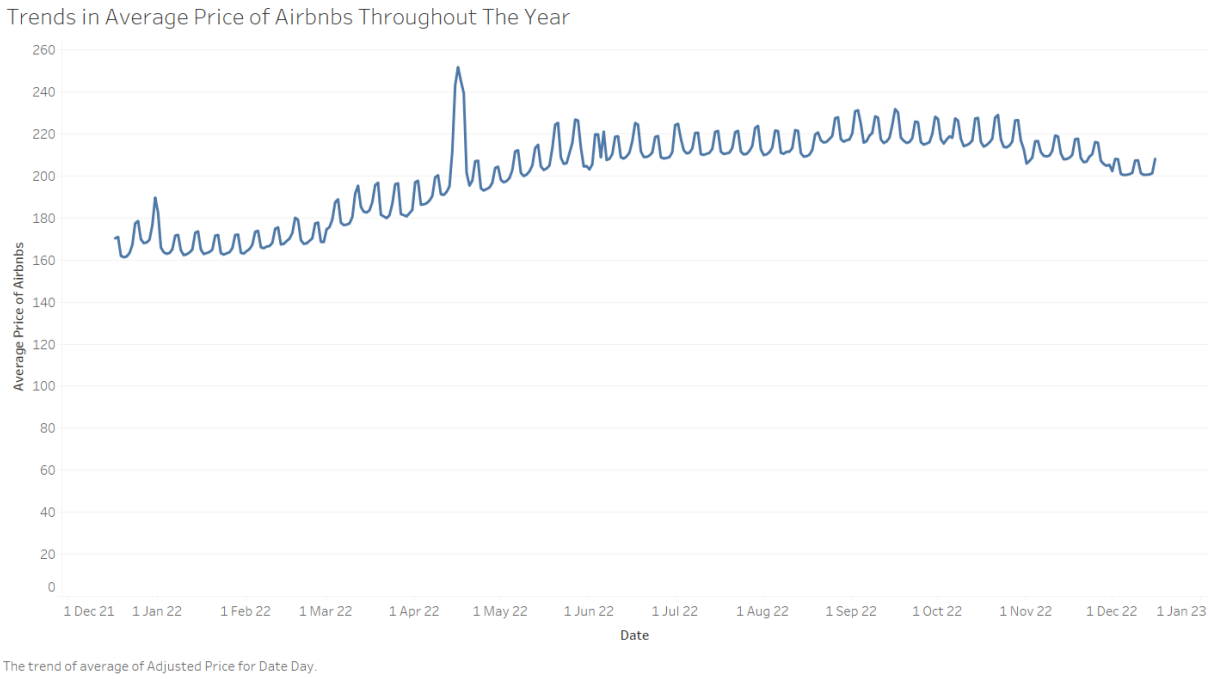
It can be observed that rooms in boutique hotels and hotels are the most expensive overall, while shared rooms in condominiums and townhouses are the least expensive, which are as expected.

8) Average ratings of Superhosts vs. Non-Superhosts



Superhosts on average have a rating higher than 4.8 while non-superhosts have an average rating of just over 4.5 which is a significant gap.

9) Trends in Average Price of Airbnbs throughout the year



There are two prominent trends we can observe above. One is the weekly increase in prices on the weekends and the other is the seasonal trend of the prices which increases during the warmer months and reduces during colder weather.

7. Modification History

Version #	Date	Author	Description
1	4/6/2022	Team	Initial Document
2	4/21/2022	Team	Added details about implementation in various sections
3	5/4/2022	Team	File Preprocessing and SSIS Implementation
4	5/4/2022	Team	Business Intelligence

8. References and Data Sources

1) Airbnb data via InsideAirbnb - http://insideairbnb.com/get-the-data/
2) US Census Data - https://www.census.gov/data/datasets