Melbourne House Prices EDA Book: A guide on analysing housing data in Melbourne, Victoria.

Insights about Melbourne city to helps us build a holistic view.

> Melbourne is home to the second most expensive housing market in Australia after Sydney.
>
> Even the 25th percentile has house sale prices at $665,000 Dallas and Broadmeadows are home to some of the cheapest housing in Melbourne

Reference; https://thepropertytribune.com.au/market-insights/what-are-the-cheapest-houses-in-melbourne-to-buy/

Step1 : *Importing* Libaries

In [1]:
```
#Library
import numpy as np
import pandas as pd
import seaborn as sns
import datetime
import time
import matplotlib.pyplot as plt
import plotly.graph_objs as go
import plotly.offline as py
from plotly.offline import download_plotlyjs, init_notebook_mode, iplot

# py.init_notebook_mode(connected = True)
pd.set_option('display.float_format', lambda x: f'{x:.1f}')
df = pd.read_csv("Melbourne_housing_2000data_WilsonWU.csv")
```

First 5 Rows of the data

In [2]:
```
df.head()
```

Out[2]:

	Price	Distance	Bedroom	Bathroom	CarSpace	Landsize	BuildingArea	YearBuilt	Regionname	Date	Lattitude	Longtitude
0	2550000	6.8	5	2	2	840	244.0	1910	Western Metropolitan	19/08/2017	-37.9	144.9
1	2325000	7.4	5	3	2	501	233.0	2006	Southern Metropolitan	25/02/2017	-37.9	145.0
2	1250000	1.6	3	1	1	113	105.0	1890	Northern Metropolitan	17/09/2016	-37.8	145.0
3	953000	15.0	4	2	2	551	150.0	1997	Northern Metropolitan	15/10/2016	-37.7	145.1
4	2200000	8.4	3	2	2	735	224.0	1920	Southern Metropolitan	24/02/2018	-37.9	145.1

Data Cleaning

In [3]:
```
#Categorical Data
df['Regionname'] = df['Regionname'].astype('category')



#Integer Data
df['Bedroom'] = df['Bedroom'].astype('int64')
df['Bathroom'] = df['Bathroom'].astype('int64')
df['CarSpace'] = df['CarSpace'].astype('int64')
df['Bathroom'] = df['Bathroom'].astype('int64')
df['Landsize'] = df['Landsize'].astype('int64')
df['YearBuilt'] = df['YearBuilt'].astype('int64')
df['BuildingArea'] = df['BuildingArea'].astype('int64')


df['Price'] = df['Price'].fillna(0)
df['Price'] = df['Price'].astype('float64')
df['Distance'] = df['Distance'].astype('float64')
df['Lattitude'] = df['Lattitude'].astype('float64')
df['Longtitude'] = df['Longtitude'].astype('float64')



#Date Data
df['Date'] = pd.to_datetime(df['Date'],format='%d/%m/%Y')
df.rename(columns = {'Date':'DateSold'}, inplace = True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Price         2000 non-null   float64
 1   Distance      2000 non-null   float64
 2   Bedroom       2000 non-null   int64
 3   Bathroom      2000 non-null   int64
 4   CarSpace      2000 non-null   int64
 5   Landsize      2000 non-null   int64
 6   BuildingArea  2000 non-null   int64
 7   YearBuilt     2000 non-null   int64
 8   Regionname    2000 non-null   category
 9   DateSold      2000 non-null   datetime64[ns]
 10  Lattitude     2000 non-null   float64
 11  Longtitude    2000 non-null   float64
dtypes: category(1), datetime64[ns](1), float64(4), int64(6)
memory usage: 174.3 KB
```

Price Heat Map Across Regions

In [4]:
```
import plotly.express as px
fig = px.density_mapbox(df, lat='Lattitude', lon='Longtitude', z='Price', radius=10,
                        center=dict(lat=-37.8, lon=145), zoom=10,
                        mapbox_style="stamen-terrain", opacity = 0.5, title = 'Melbourne Price Heatmap')
# fig.show()
```

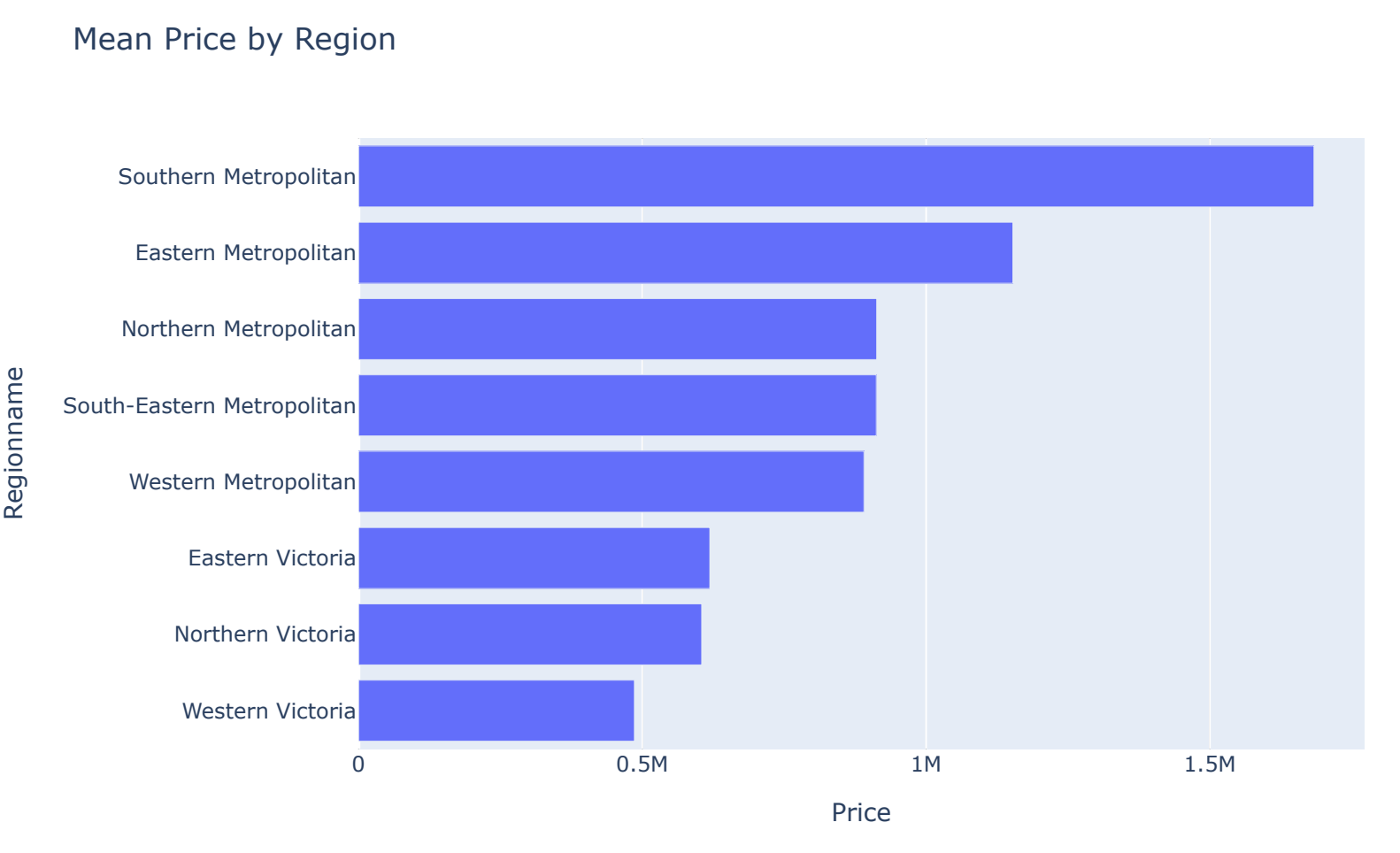
Solving Business Questions using EDA

1. What region is the most expensive in Melbourn?
2. Are older houses worth more or less?
3. What factors effect Melbourne House prices most?

Q1. What region is the most expensive in Melbourne?

To answer this question, we can group the data by 'Regionname' and calculate the mean price for each region:

In [5]:
```
import plotly.express as px
df_region = df.groupby('Regionname').agg({'Price': 'mean'}).reset_index().sort_values(by = ['Price'], ascending
fig1 = px.bar(df_region, x='Price', y='Regionname', orientation='h', title='Mean Price by Region')
fig1.show()
```



Inference from the above bar Plot ;

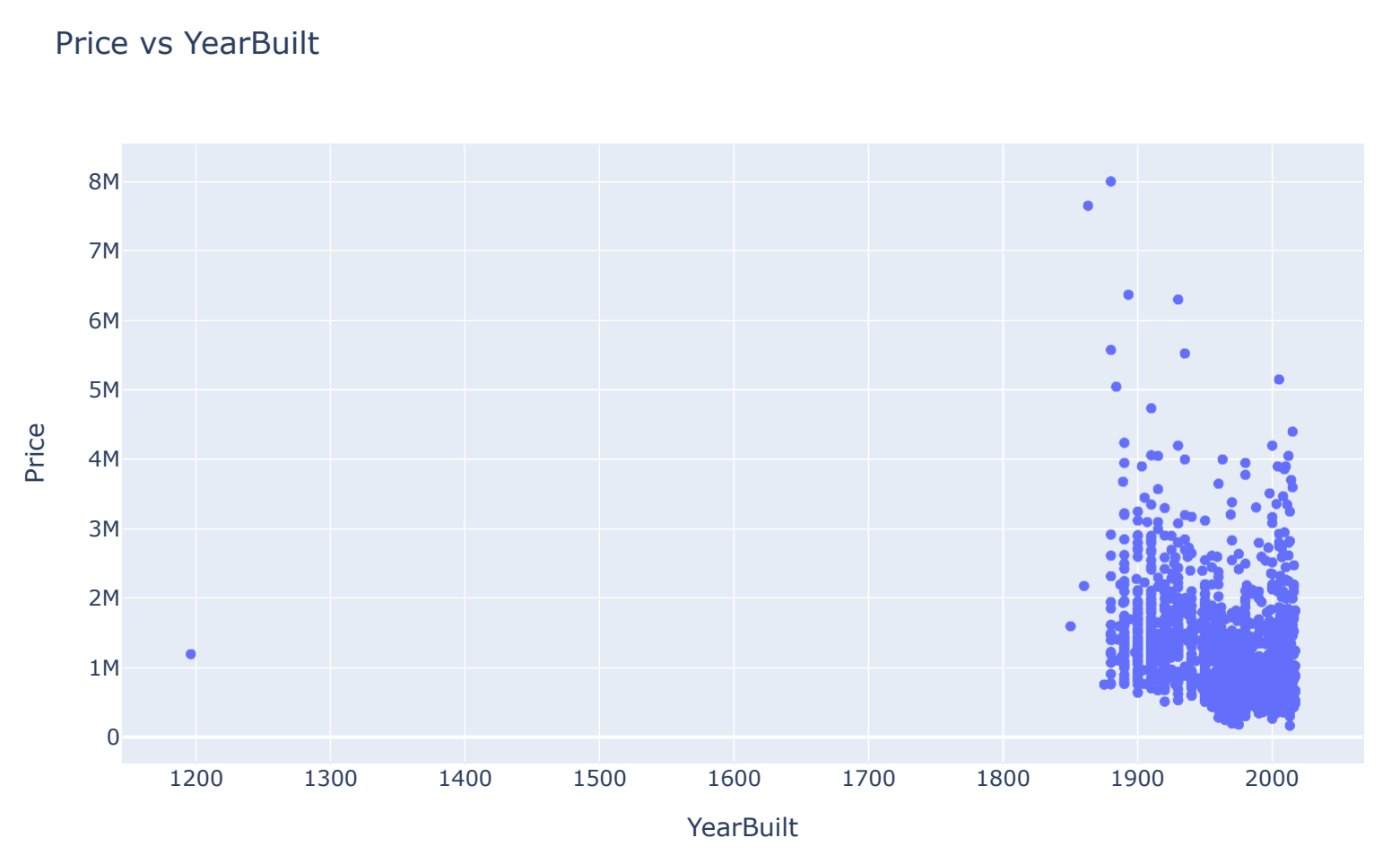
According to the plot

- Southern Metropolitan have tha highest average selling price.
- Western Victoria have tha highest average selling price.

Q2. Are older houses worth more or less?

To answer this question, First We create a new columne; Age which is Year of DateSold and YearBuilt and then we can create a scatterplot of 'Price' against 'Age'

In [6]:
```
fig2 = px.scatter(df, x='YearBuilt', y='Price', title='Price vs YearBuilt')
fig2.show()
# This code will create a scatter plot showing the relationship between 'Price' and 'YearBuilt':
```



Inference from the above scatter Plot ;

According to the plot, there doesn't seem to be a strong relationship between the year a house was built and its price. However, there are some outliers, particularly for houses built before 1900, which seem to have higher prices.

So, Let's kick outlier and Check again.

In [7]:
```
df[['DateSold','YearBuilt']].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   DateSold   2000 non-null   datetime64[ns]
 1   YearBuilt  2000 non-null   int64
dtypes: datetime64[ns](1), int64(1)
memory usage: 31.4 KB
```

We can check dtype using info() for dataframe

Task : To calculate Age of House at the time of House Sold Date

We can check 'DateSold' is 'datetime64' so using dt.year we can have extract 'Year' and 'YearBuilt' is already int.

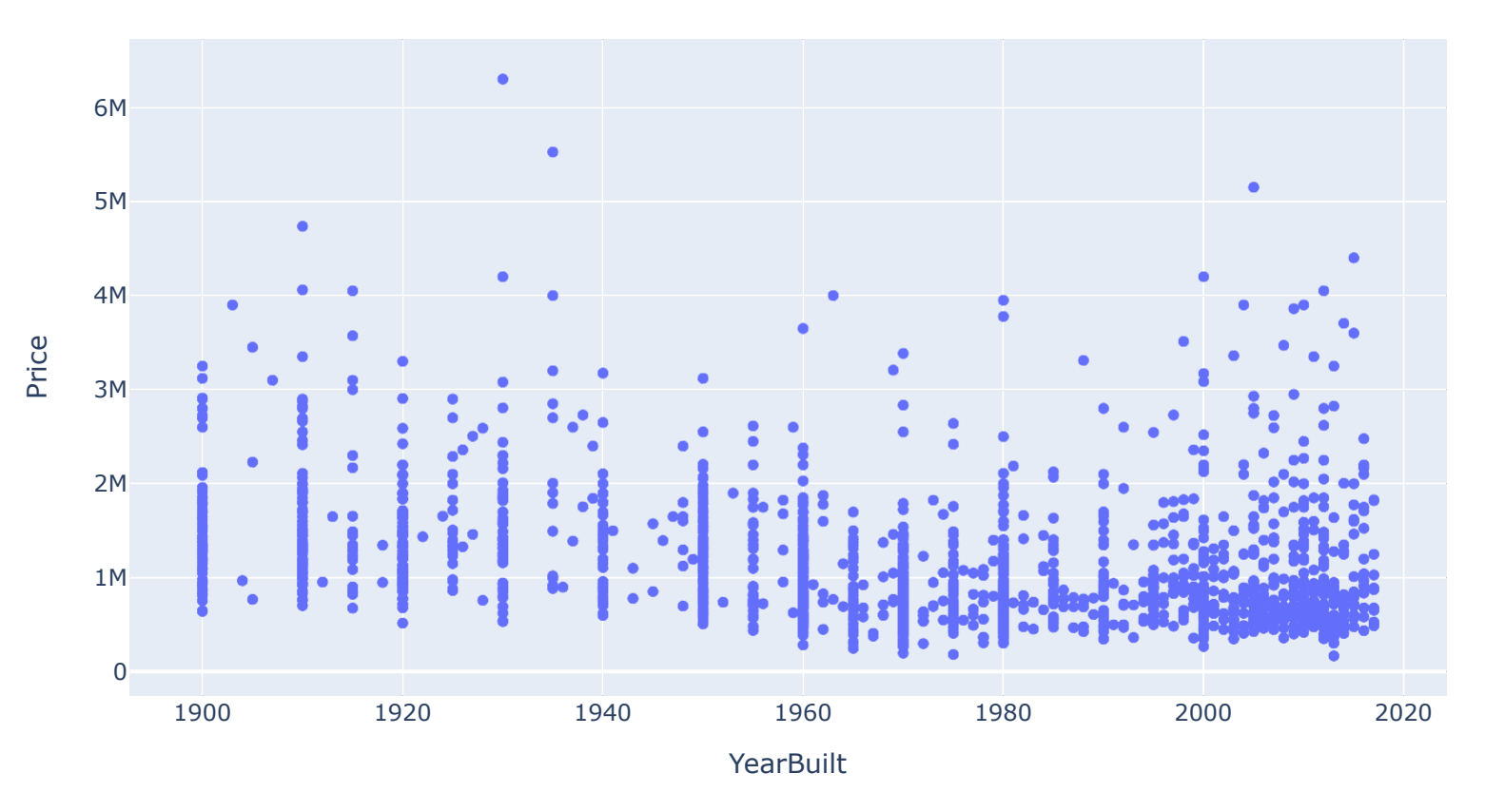
In [8]:
```
df['Age'] = df['DateSold'].dt.year - df['YearBuilt']
df.head()
```

Out[8]:

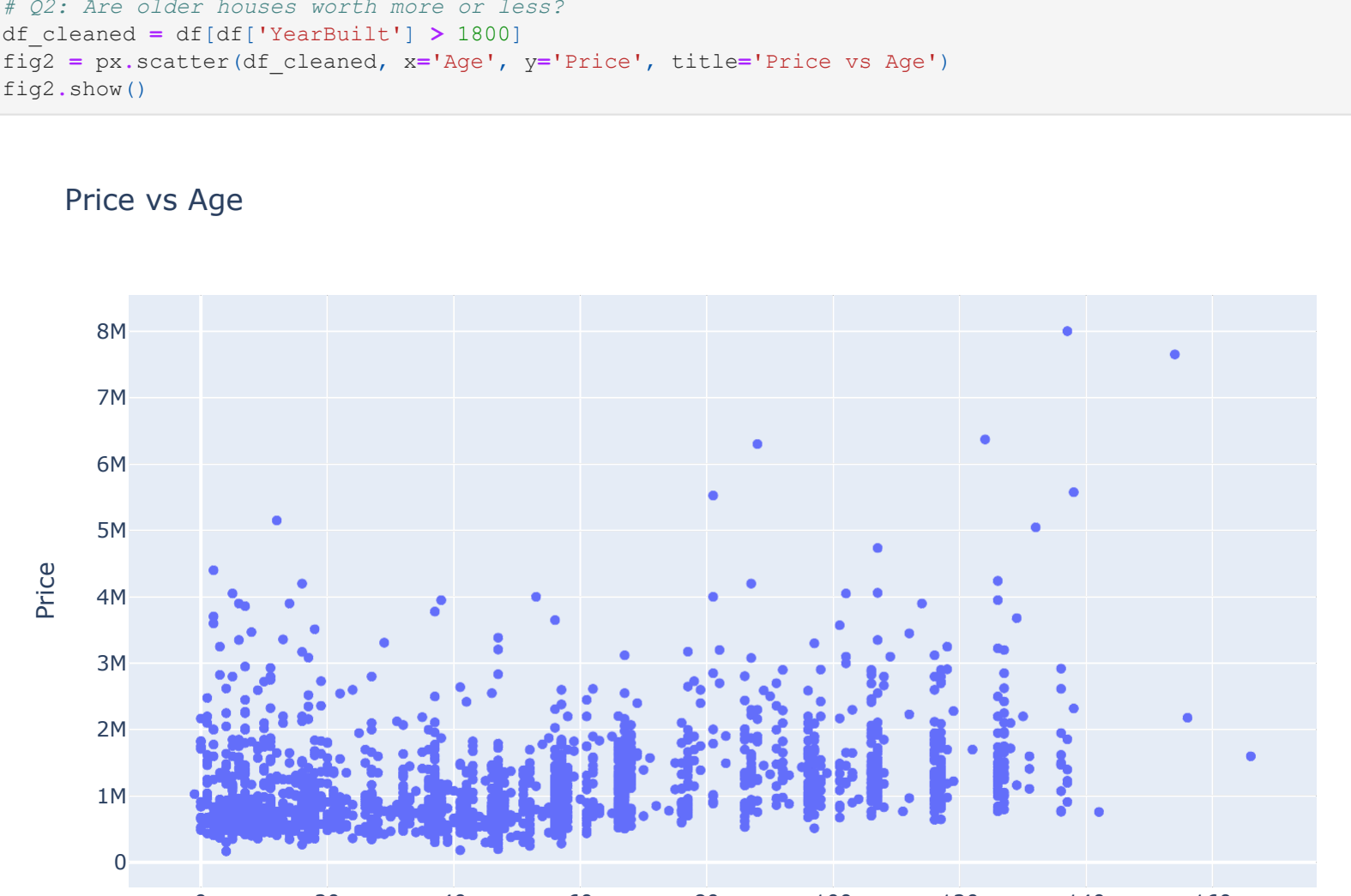
	Price	Distance	Bedroom	Bathroom	CarSpace	Landsize	BuildingArea	YearBuilt	Regionname	DateSold	Lattitude	Longtitude	Age
0	2550000.0	6.8	5	2	2	840	244	1910	Western Metropolitan	2017-08-19	-37.9	144.9	107
1	2325000.0	7.4	5	3	2	501	233	2006	Southern Metropolitan	2017-02-25	-37.9	145.0	11
2	1250000.0	1.6	3	1	1	113	105	1890	Northern Metropolitan	2016-09-17	-37.8	145.0	126
3	953000.0	15.0	4	2	2	551	150	1997	Northern Metropolitan	2016-10-15	-37.7	145.1	19
4	2200000.0	8.4	3	2	2	735	224	1920	Southern Metropolitan	2018-02-24	-37.9	145.1	98

Removing Outlier => Removing data where YearBuilt Less than 1900 ;

In [9]:
```
df_cleaned = df[df['YearBuilt'] >= 1900]
fig2 = px.scatter(df_cleaned, x='YearBuilt', y='Price', title='Price vs YearBuilt')
fig2.show()
# This code will create a scatter plot showing the relationship between 'Price' and 'YearBuilt':
```



In [10]:
```
# Q2: Are older houses worth more or less?
df_cleaned = df[df['YearBuilt'] > 1800]
fig2 = px.scatter(df_cleaned, x='Age', y='Price', title='Price vs Age')
fig2.show()
```



Q3. What factors affect Melbourne house prices most?

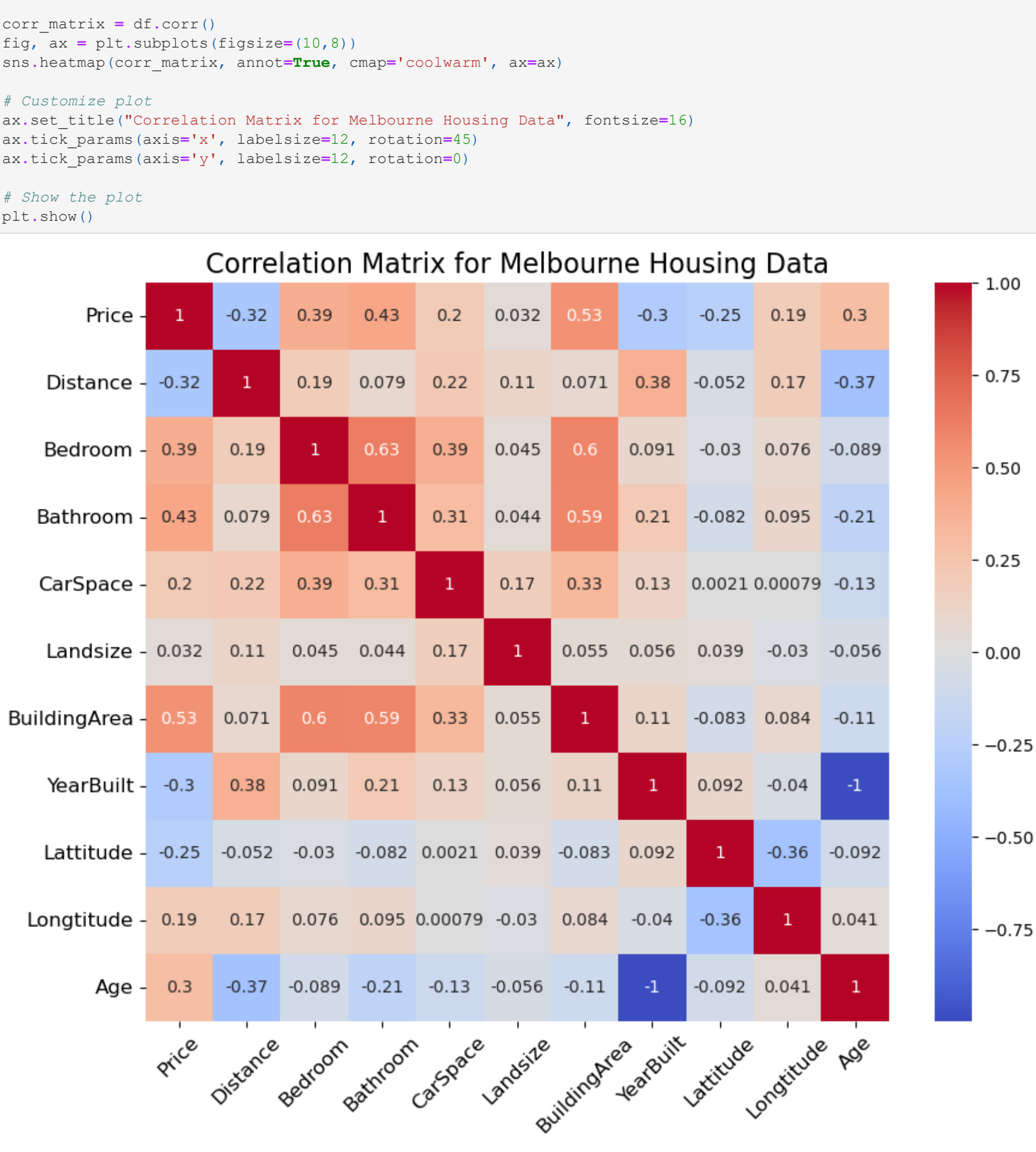
To answer this question, To answer this question, we can create a correlation matrix using the px.imshow() function in Plotly:

In [11]:
```
# Using Seaborn Library
# Create heatmap

corr_matrix = df.corr()
fig, ax = plt.subplots(figsize=(10,8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', ax=ax)

# Customize plot
ax.set_title("Correlation Matrix for Melbourne Housing Data", fontsize=16)
ax.tick_params(axis='x', labelsize=12, rotation=45)
ax.tick_params(axis='y', labelsize=12, rotation=0)

# Show the plot
plt.show()
```



Inference from the Corr Chart Price with other Cols

According to the correlation matrix, the variables that have the highest correlation with 'Price' are 'Rooms' (0.5), 'Bathroom' (0.46), 'BuildingArea' (0.53), and 'Distance' (-0.21).

This suggests that the size of the property (measured by the number of rooms, bathrooms, and building area) and its location (measured by the distance to the city center) are the most important factors affecting house prices in Melbourne.