**Written Analysis Report**

**Abstract:**

The real estate market is constantly changing and characterized by an incredibly multivariate nature. When gleaning insights into this market it’s important to take a granular approach. As such, we decided to look specifically at “input” variables and their respective impacts on listing price, the variable of time and its influence on market trends as well as comparing key housing metrics. Combining the insights gained from each approach resulted in the following conclusions within our selection criteria.

**Data Cleaning:**

The steps in the cleaning process were:

1. Our primary csv for real estate data contained more than 2M datapoints and was significantly larger than GitHub allowed (170.57 MB). As such, we trimmed down our dataset logically by removing unnecessary columns (Brokered By & Street) and only including columns of interest such as status, price, bed, bath, acre lot, city, state, zip code, house size and previous sold date.
2. Removed U.S. Territories (Puerto Rico & Virgin Islands) to only include states within our analysis.
3. We wanted to analyze housing data from different regions of the country to determine how impactful different variables were on the housing market in these regions. Therefore, we chose Washington, Wisconsin and Pennsylvania as a cross section of the Unites States (West, Mid-West and East).
4. To make this analysis more impactful, we decided to filter our dataset to represent listings with 5 bedrooms or less as we felt it was appropriate for the average real estate market participant (sellers, buyers, investors, developers and real estate professionals).
5. As the real estate market is highly variable and heavily dependent on time, we determined an approximate 10-year range would be suitable to conceptualize "current" market trends. To do so, it was important that we converted the raw data within the "previous sold date" column to "datetime" format.
6. For the final stage of our data cleaning process, we determined and eliminated outliers via interquartile range calculations. Additionally, we eliminated listings with a price of less than 5,000 dollars. This was executed to remove inaccurate listing due to data entry error (such as listing with a $1.00 price, listings with information entered into the wrong field, etc.).

**Regression Analysis:**

Within our “Regression Analysis” Jupyter notebook we explored and cleaned our U.S real estate listing dataset to retrieve key metrics about the data as well as performing several regressions on specific input variables (inclusive of two metrics sourced via two different APIs) to determine their respective impacts on listing prices.

1. Prior to determining any correlations, we needed to clean our data, remove any potential outliers and filter datapoints to represent our specific selection criteria. Please refer to the “Data Cleaning” section above for specific steps surrounding the cleansing process, outlier detection and elimination as well as clear definition of our selection criteria
2. Once our data was cleaned, filtered and ready for processing we first wanted to calculate and display the measures of central tendency as well as maximums and minimums for pricing for each individual state. As represented by the associated visualization, we can see that Washington displays significantly higher price metrics for mean, median, maximum and minimum.

**Wisconsin:** Mean = $275,966.59, Median = $250,000, Mode = $149,900 (x215), Max = $694,900, Min = $9,900

**Washington:** Mean = $631,944.84, Median = $559,000, Mode = $425,000 (x145), Max = $1,539,000 Min = $47,900

**Pennsylvania:** Mean = $272,983.45, Median = $249,000, Mode = $199,900 (x517), Max = $719,900, Min = $6,500

1. After establishing the initial measures above, we subsequently performed bivariate linear regressions for each state’s price outcome vs. the following input variables: Number of Bedrooms, Number of Bathrooms, Land Size (Acres), Living Space (Square Feet), City Population and City Maximum Temperature. Each city’s maximum temperature (as well as unique Latitude and Longitude) value metric was retrieved via the OpenWeather API. We established a unique list of city names for each state and ran a get request for each city and parsed the JSON response to return unique latitude, longitude and maximum temperature. During this request we included a progress indicator and try/except code language which returned a message detailing all cities that were unable to be processed by the API. Once obtained, the data was processed into a dataframe ahead of the second API request.

The second API request was facilitated via the Geoapify API in order to retrieve a population number for each unique city. Again, within this call loop we included try/except coding language as well as a progress indicator that listed whether or not each city had success within the Geoapify database to return the JSON parsed value. These population values were added to our previously established city dataframe within the introduced “Population” column.

Once the API requests were completed, we merged our retrieved city data with our master cleaned datasets for each state, dropped any datapoints where we were unable to populate the max temperature/population values so we could run subsequent regression analysis on these new variables. The results for each state’s regressions on price are listed below inclusive of the linear equation for predictive analysis:

**Wisconsin:**

Bedrooms vs. Price: r-value = 0.33, r-squared value = 0.11, linear equation = (y = 49046.24x + 116860.16)

Bathrooms vs. Price: r-value = 0.63, r-squared value = 0.39, linear equation = (y = 94486.79x + 72091.75)

Land Size (Acres) vs. Price: r-value = 0.11, r-squared value = 0.01, linear equation = (y = 1492.26x + 273696.30)

Living Space (Sq. Ft.) vs. Price: r-value = 0.64, r-squared value = 0.41, linear equation = (y = 110.12x + 66893.54)

Max Temperature vs. Price: r-value = 0.25, r-squared value = 0.06, linear equation = (y = 5005.14x + -48860.74)

City Population vs. Price: r-value = -0.01, r-squared value = 0.00003, linear equation = (y = -0.02x + 263382.80)

**Washington:**

Bedrooms vs. Price: r-value = 0.35, r-squared value = 0.12 , linear equation = (y = 103465x + 288313.26)

Bathrooms vs. Price: r-value = 0.49, r-squared value = 0.24, linear equation = (y = 171718.28x + 228060.54)

Land Size (Acres) vs. Price: r-value = 0.11, r-squared value = 0.01 , linear equation = (y = 8263.84x + 623692.41)

Living Space (Sq. Ft.) vs. Price: r-value = 0.57, r-squared value = 0.33, linear equation = (y = 194.78x + 238336.04)

Max Temperature vs. Price: r-value = -0.20, r-squared value = 0.04, linear equation = (y = -7136.01x + 971264.07)

City Population vs. Price: r-value = 0.03, r-squared value = 0.001, linear equation = (y = 0.55x + 512287.92)

**Pennsylvania:**

Bedrooms vs. Price: r-value = 0.29, r-squared value = 0.09 , linear equation = (y = 45466.12x + 123088.37)

Bathrooms vs. Price: r-value = 0.62, r-squared value = 0.39, linear equation = (y = 98380.78x + 56870.90)

Land Size (Acres) vs. Price: r-value = 0.01, r-squared value = 0.0001 , linear equation = (y = 68.79x + 272924.55)

Living Space (Sq. Ft.) vs. Price: r-value = 0.06, r-squared value = 0.004, linear equation = (y = 1.15x + 270856.06)

Max Temperature vs. Price: r-value = -0.20, r-squared value = 0.04, linear equation = (y = -7771.23x + 825580.85)

City Population vs. Price: r-value = 0.03, r-squared value = 0.0008, linear equation = (y = 0.05x + 273039.28)

1. After conducting the six regression models for each state we loaded our results into a single visualization that displays each variables respective r-value/correlation on price for each state. This allowed us to visually compare the strength of each correlation and determine if there were any discrepancies between our three states. While Wisconsin’s and Washington’s two strongest correlations on price was living space followed by number of bathrooms respectively, Pennsylvania’s two strongest correlations on price were number of bathrooms first with number of bedrooms second. Another interesting conclusion from our analysis and this visualization specifically, was that there seems to be a sizeable discrepancy when it comes to Pennsylvania’s living space and land size correlations on price. The land size correlation was less than the two other states and significantly less when it comes to its living space correlation on price.

**Investment Analysis:**

Within our “Investment Analysis” Jupyter notebook we wanted to take a closer look at some of our dataset’s key metrics, how time influenced the real estate market within our three states and the associated selection criteria as well as any conclusions we could potentially put forth to a market participant such as an investor or real estate developer.

1. First, it was important for us to establish a clear dichotomy for our listing status between homes “for sale” and homes “sold”. As evidenced by these pie charts, we can see that the distribution in Wisconsin is represented by 55.1% of listings falling into the “sold” category and 44.9% within the “for sale” category. This breakdown is quite similar to that of Pennsylvania with 55.2% of listings being “sold” and 44.8% “for sale”. Interestingly, we see a very different distribution within the state of Washington with a miniscule 2.7% “sold” and a whopping 97.3% “for sale”.

Further representing this trend is another pie chart that depicts the overall comparison of number of homes sold respective to each state. Pennsylvania greatly outperforms its counterparts with 75.7% of all the “sold” homes within our dataset. Coming in second is Wisconsin with 23.3% followed by Washington at only 1%. This data clearly shows the density of the Pennsylvania real estate market and conveys its primacy as it relates to inventory velocity/turnover within our selection criteria.

1. To better understand how time effects the overall number of houses sold for each state, we’ve created a line chart which visually represents an increase in the total number of houses sold for each state from 2021 to 2022. Adding to the validity of Pennsylvania’s high degree of inventory velocity, we can see the rate of increase or slope for the state is significantly steeper than that of Wisconsin and Washington.
2. Our next visualization continues to represent the trends for each state’s “sold” vs. “for sale” statistics specifically within the lens of total number of houses and the average sale price. We can very clearly see that our data experiences a significant change within 2020. According to the “Houses Sold” line depicted in blue for each graph, we see a remarkable increase in the total number of houses sold across all states in 2020 and thereafter. As an increase in houses sold represents an increase in market demand and subsequent decrease in available inventory, we can see the same distinct uptick visually in terms of average price for homes sold as well (right side graphs). It is important to note that our dataset unfortunately doesn’t contain adequate “sold” information for listings prior to 2020 and as such this instance is listed within the “Limitations” section of our project scope document underneath “Dataset Integrity”.

Another interesting relationship regarding market inventory trends within our dataset can be visualized by the delta or difference between the red “for sale” trendline in direct comparison to the blue “sold” trendline. When looking at this relationship on the left side series of graphs for total number of houses sold, all three states experience a smaller delta between these two trendlines post 2020. This shows that the changing relationship between “sold” and “for sale” metrics impacting the housing market’s supply and demand. Historically, for all three states we can see a consistent delta between the number of homes “for sale” and homes “sold” and a subsequent shrinking occurring around 2020 (homes for sale out numbering homes actually sold). In the cases of Wisconsin and Pennsylvania, this impact even results in the number of homes “sold” greatly outweighing the number of homes “for sale”.

On the average price side (right side series of graphs), we can see from our 2020 flashpoint onward there’s a similar effect which results in “for sale” average prices better aligning to their counterpart of actual “sold” average prices. In other words, the discrepancy between average listing/for sale prices and the average price for which homes are being sold decreases. For Pennsylvania and Wisconsin, so much so that there was a point where the average price for homes sold was higher than the average listing price.

In order to show these relationships in a collective manner for our dataset, we lastly created a visualization showing the “for sale” and “sold” trendlines for each state on a single graph for number of houses sold and average price. This further demonstrates the impact of time on the real estate market and represents that impact for each state allowing comparison.

1. Finally, pulling from a number of our conclusions we wanted to put forth some implications which we could potentially provide to a market participant. In an effort to do so and in conjunction with our above analysis, we created a number of additional visualizations that display the total dollar amount sum of all houses sold for each state, the average price for a home sold within each state as well as state specific averages for price per square foot, number of bathrooms, number of bedrooms and square feet of living space. We also included a bar graph to visualize and compare each state’s standard deviation with regards to pricing to assess “riskiness”. As you can see, the first graph shows the highest volume performing state was Pennsylvania with $5,849,285,276, the lowest volume performing state was Washington with $123,429,533 and Wisconsin in-between the two with a total dollar amount sum of $1,743,383,205.

When looking at the other graphs, our calculations display Washington maintains the number 1 spot across all averages including the highest average price for a home sold at $422,703.88, followed by Pennsylvania at $264,434.23 which was very slightly ahead of Wisconsin with a value of $256,191.51. Washington also has the highest price per square foot at $330.62 (Pennsylvania at $155.51 & Wisconsin at $150.25), the highest average number of bathrooms at 2.35 (Pennsylvania at 2.20 & Wisconsin at 2.16), the highest average number of bedrooms at 3.32 (Pennsylvania at 3.30 & Wisconsin at 3.24), and finally the highest average living space with 2,020.75 square feet (Pennsylvania at 1,851.50 & Wisconsin at 1,898.56). The listings within the state of Washington also experience the highest pricing standard deviation with a value of $293,409.39 followed by Pennsylvania at $147,015.86 very slightly ahead of Wisconsin with a value of $138,374.87.

Given these final calculations/visualizations along with the entirety of our analysis, we would suggest a few key points to consider for a potential real estate developer or investor looking to put money into the real estate market within one of our three states. First, our data suggests it’s a solid time to be investing in the real estate market as evidenced by a strong uptick in demand mirrored by an increase in average pricing across the board. Although every key metric (pricing measures of central tendency, maximum/minimum, average price of a home sold, average price per square foot, average living space, average number of bedrooms/bathrooms) is clearly in favor of the Washington real estate market, the poor inventory velocity within the state coupled with the highest standard deviation regarding price could represent a “riskier” investment. While the return from a single house sold in Washington would indeed outweigh the return experienced in Wisconsin or Pennsylvania, the quantity of houses being sold/highest rate of increase and in-turn the total sum of dollars is maximized within Pennsylvania. As such, if the investment strategy revolved around casting a “wide net” with multiple properties, we’d suggest Pennsylvania would represent the best option within our three states. Furthermore, as concluded via our regression analysis, within Pennsylvania one should place more emphasis on a property’s number of bathrooms first, followed by the number of bedrooms above other characteristics.

Outside of the real estate industry specifically, policy makers can also utilize this analysis to inform decisions surrounding future growth, supply and demand, specific legislation/programs to adequately influence their local real estate market.