Good Morning,

My name is Justin. I am a new hire on the data science team. I was asked to brief you on the results of our most recent prediction project.

We were tasked with collecting 2017 data on Single Family Properties from the Codeup SQL database servers. Clean it, prepare it, analyze and explore it so that we could select features and models to process through machine learning with these specific goals:

- Predict property tax assessed values of Single Family Properties
- Outperform existing logerror model
- Recommend improvements for a more accurate model

- Define relevant fips codes for our data, since Maggie seems to have misplaced the cheat sheet.

ACQUISITION:

- The data was initially pulled on 15-NOV-2022.
- The initial DataFrame contained 52,441 records with 69 features (69 columns and 52,441 rows) before cleaning & preparation.
- Each row represents a Single Family Property record with a Tax Asessment date within 2017.
- Each column represents a feature provided by Zillow or an informational element about the Property.

Summary of Data Cleansing

- Cleaning the data resulted in less than 10% overall record loss

- **There were** 39 features each containing over 30% NaN that were dropped; **resulting in no record loss.**
- **There were** 1,768 records containing NaN across 13 features that were dropped; **resulting in only 3% record loss.**
- **OUTLIERS:** there were aproximately 3,000 outliers that were filtered out in an attempt to more accurately align with realistic expectations of a Single Family Residence; **resulting in less than a 6% decrease in overall records.**
- **No data was imputed**

There is a Data Dictionary for your reference

Here's a sneak peek at the data

It was split into train, validate, test at roughly a 60/20/20 ratio

There are over 28,000 records in our training data consisting of 18 features.

- Exploration of the data was conducted using various Correlation Heat Maps, Plot Variable Pairs, Categorical Plots, and many other graph and chart displays to visualize Relationships between independent features and the target as well as their relationships to eachother.
- Each of the three selected features were tested for a relationship with our target of Tax Assesed Value.
    1. Bedrooms
    2. Bathrooms
    3. Property Squarefeet

- All three independent features showed a significant relationship with the target feature.
- Three statistical tests were used to test these hypothesis.
  1. T-Test
  2. Pearson's R
  3. KAI-Squared

Here we see a summary each Question, the Hypothesis and the results of the statistical test, followed by a visualization of the data.

[Read the Question]

As you can see all of these relationships appear to be fairly linear trending in an upward pattern.

[READ TAKEAWAYS]

We used Multiple Regression + RFE along with correlation heat maps to decide on the three features with the strongest relationship to our targets they were Bedroom, Bathroom, and SQFT

Moving into modeling,
Baseline and logerror are our evaluation metrics

We used 6 different regression models to explore the data. Below are the results of those models.

`2nd Degree Polynomial` is the best model and will likely continue to perform well above Baseline and logerror on the Test data.

[READ CONCLUSION]