**Project Report: Singapore Fin-tech**

**1. Introduction**

**1.1 Background**

This is the project report on **Property Tax** which is part of Singapore Fin-tech project. We can get valuable insights from this data how property tax collection varies based on different property types through the years.

**1.2 Problem Statement**

Analyze the given data and present statistical insights such as Mean, Median, Standard Deviation, Skewness, Kurtosis, Percentiles, Correlation Coefficient, Chi-square Test, Outlier Detection & Statistical graphs.

Use ML techniques wherever required.

**1.3 Aim & Objectives**

**Aim:** To apply statistical analysis to derive statistical insights and apply ML model to predict the property tax collection. Perform Chi-square test to test the accuracy of ML model.

**Objectives:**

* Clean and pre-process the Property tax dataset.
* Apply statistical techniques to gain valuable insights.
* Explore the distribution and summary statistics of overall data.
* Visualize the distribution property tax by property type.

**3. Data Acquisition and Preprocessing**

This chapter will detail the data acquisition and preprocessing steps undertaken in the project. It will cover the following:

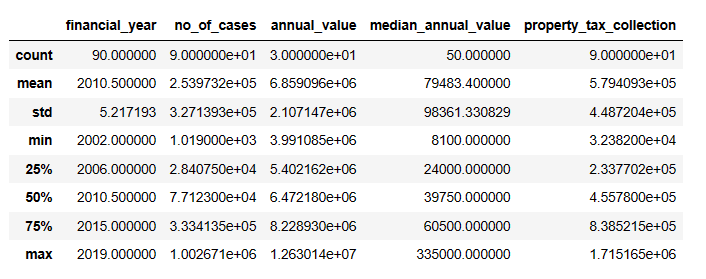
* **Data source:** -----------------.
* **Data loading:** Data Loaded using Python’s Pandas library in Jupyter Notebook.
* **Data cleaning:** Found presence of several null values hence filled the null values with median of the data as per type of property type.

**4. Data Analysis and Exploration**

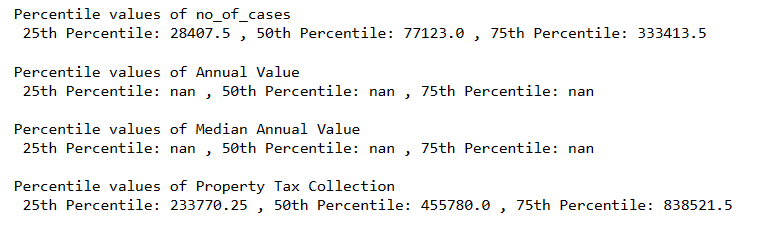
Below are the statistical findings:

* **Descriptive statistics:** Below are the descriptive statistical findings:

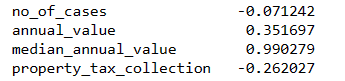
Mean, Median, Standard Deviation



Percentile

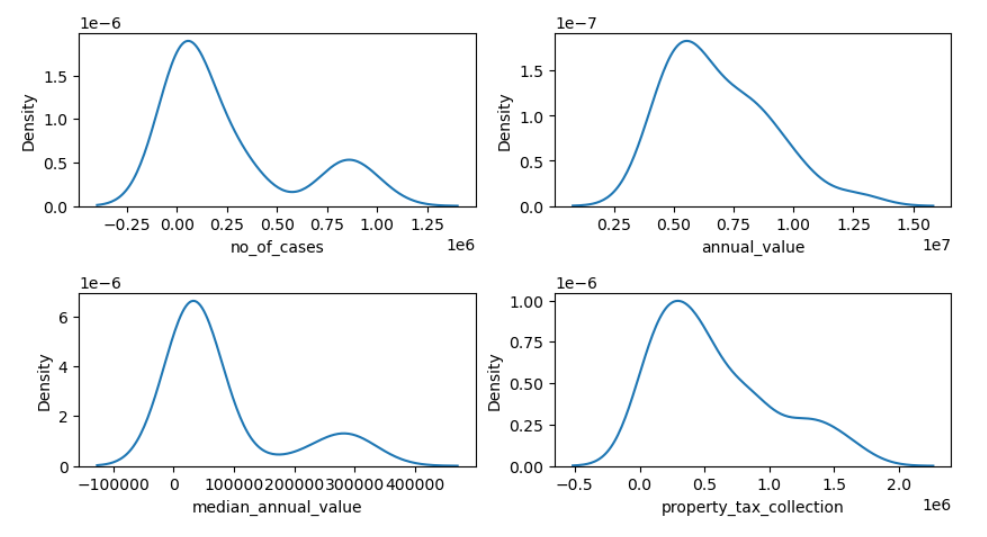


Kurtosis



* **Data visualization:**

KDE Plot

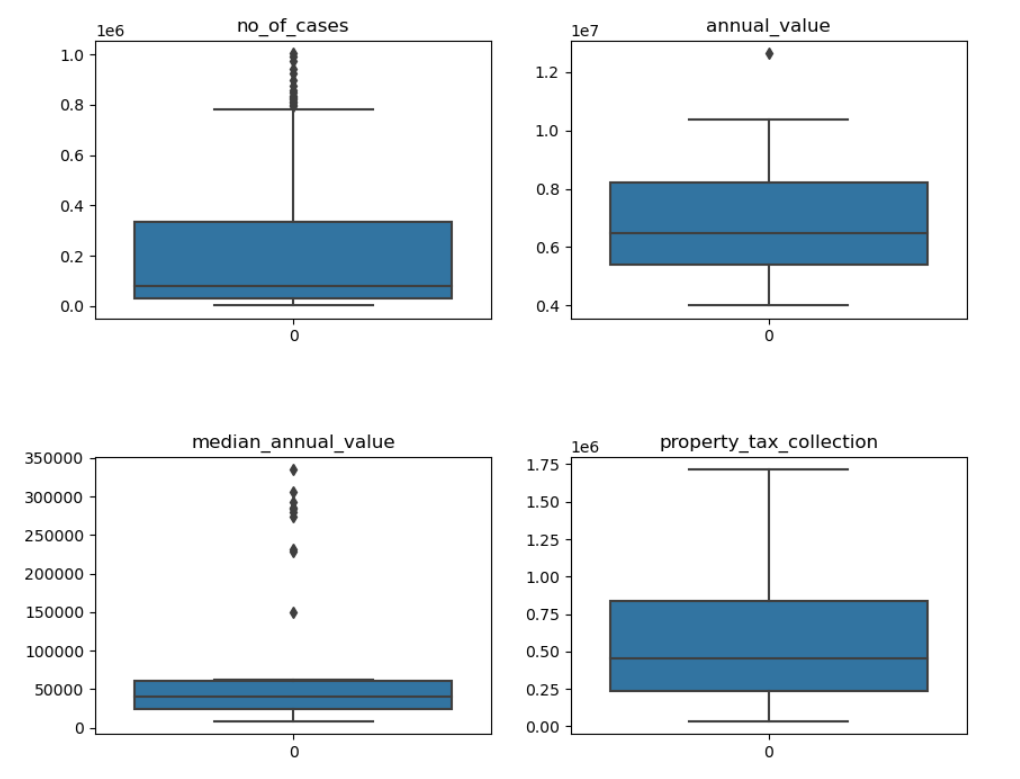


* **Statistical tests:** Conducted Chi-square test after applying ML model.

**5. Findings and Discussion**

Below are key findings:

* **Key insights:**
  + As per KDE plot we it is observed that the data is skewed towards right side.
  + Data contains few outliers. Identified with Boxplot.



* **Discussion:** A chi square value of 0.0 suggests that the ML model worked as per expectations and predicting the correct values.

**6. Work Methodology**

**6.1 Introduction**

This project involved a multi-step methodology, including data loading, pre-processing, application of statistical functions, detailed data analysis, application ML model and statistical test.

**6.2 Research Approach**

* **Data Loading:** Data Loaded using Python’s Pandas library in Jupyter Notebook.
* **Data Preprocessing:** Filled null values using Median values according to the property type. Before applying ML model the categorical data was converted into numerical data using pd.getdummies() function of pandas library.

Encoded categorical variables ('property\_type') using one-hot encoding, which is essential for machine learning models to interpret categorical data.

**Train-Test Split:** Performed a train-test split, with 80% of the data used for training the model and 20% for testing its performance. This is a standard practice to evaluate how well your model generalizes to unseen data.

Model Training:

We've chosen a Decision Tree Regressor as our machine learning model and trained it on the training data. Decision trees are capable of capturing complex relationships in data and are suitable for regression tasks like predicting property tax amounts.

**Model Evaluation:**

Evaluated our model's performance using the R-squared (coefficient of determination) and Mean Squared Error (MSE) metrics.

**R-squared :** Measures the proportion of variance in the target variable explained by the model

**MSE :** Quantifies the average squared difference between predicted and actual values.

It's important to interpret these metrics in the context of the problem domain to assess the model's effectiveness.

**Feature Scaling:**

Applied StandardScaler to standardize both input features and the target variable. Standardization is beneficial for algorithms that rely on distance-based calculations or gradient descent optimization, ensuring that features are on a similar scale.

**Cross-Validation:**

Performed cross-validation with cv=10 to assess the model's performance robustness. Cross-validation provides a more reliable estimate of the model's performance by evaluating it on multiple subsets of the data.

**Model Optimization:**

Fitted the Decision Tree Regressor model on the entire dataset post-scaling. This step is crucial for deploying the model for making predictions on new, unseen data.

**Final Evaluation:**

Evaluated the optimized model's performance using the same metrics (MSE and R-squared) on the entire dataset. This step provides a comprehensive assessment of the model's effectiveness.

* **Data Analysis:** To get the detailed statistical insights, visualization & apply ML model used Python’s libraries such as pandas, seaborn, matplotlib and sklearn.

**7. Conclusion**

The analysis of Property Tax Collection data provides valuable insights into Property Type distribution, highlighting key factors affecting values, help to find number of cases, and annual values of properties.

As per Chi-square test ML model applied on the data is predicting the correct values hence can be used to predict the property tax collection.

**8. References**

* [info@codons.in](mailto:info@codons.in)
* [www.codons.in](http://www.codons.in)