```
## E COMMERCE CUSTOMER SERVICE SATISFACTION
```

Target Company:
Shopzilla, now known as Connexity, is a company focused on online retail and ecommerce. It provides a platform for online shopping, connecting shoppers with retailers and offering a wide range of products from electronics and clothing to home and garden items. The company primarily serves the ecommerce industry. It was founded in 1996 and is based in Los Angeles, California.
Shopzilla operates a portfolio of shopping web sites, and began as a comparison shopping website. Overall stats for the company - 80.3k site visits from last month, annual revenue of about 100 - 200 million

Dataset Description:
The dataset captures customer satisfaction scores, along with multiple variables involving handling customer queries and disputes, for a one-month period at Shopzilla .

Problem statement: How does the different features of an item and the nature of the interaction with the client affect the ultimate satisfaction rating of the customer in ECommerce shopping

Significance of problem?
Customer ratings are crucial to the success of a business. Psychologically, we place value on the opinions and behaviours of others and often make decisions based on other people's decisions. Customers see them as a personal testimony. It lets us imagine the experience of a service, a process of "affective forecasting" that impacts decision making.

This is especially the case in online platforms like shopzilla as there are no physical products or shops to see, and brand reputation is a major factor in determining why people should use the platform. Good customer satisfaction score makes a platform appear more trustworthy, boosting overall brand reputation.

Intro to the dataset

Unique id	Unique identifier for each record
Channel name	Name of the customer service channel

Category	Category of the interaction
Sub-category	Subcategory of the interaction
Customer Remarks	Feedback provided by the customer
Order id	Identifier for the order associated with the interaction
Order date time	Date and time of the order
Issue reported at	Timestamp when the issue was reported
Issue responded	Timestamp when the issue was responded to
Survey response date	Date of the customer survey response
Customer city	City of the customer
Product category	Category of the product
Item price	Price of the item
Connected handling time	Time taken to handle the interaction
Agent name	Name of the customer service agent
Supervisor	Name of the supervisor
Manager	Name of the manager
Tenure Bucket	Bucket categorizing agent tenure
Agent Shift	Shift timing of the agent

CSAT Score	Customer Satisfaction (CSAT) score

DATA CLEANING AND EXPLORATION

CLEANING THE DATASET

We have selected columns denoting ‘category','Issue_reported at','issue_responded', 'Tenure Bucket', 'Agent Shift', 'CSAT Score','channel_name','Sub-category','Customer_City','Product_category','Item_price','Agent_name','Supervisor','Manager'

And out of these we would be selecting the columns that are relevant in evaluating CSAT score.

Checking for missing values

Now we will be checking for missing values among the columns we have selected.
We have decided to choose those columns that do not have many missing values as it would be difficult and not beneficial to do exploratory analysis on variables with a lot of missing values.

Dropping unnecessary columns

The variables that we dropped were the ‘Item_price’, ‘Customer_City’ and ‘Product_Category’. Now we will be putting the variables with few missing values under the cleaned dataset.

Creating a new variable

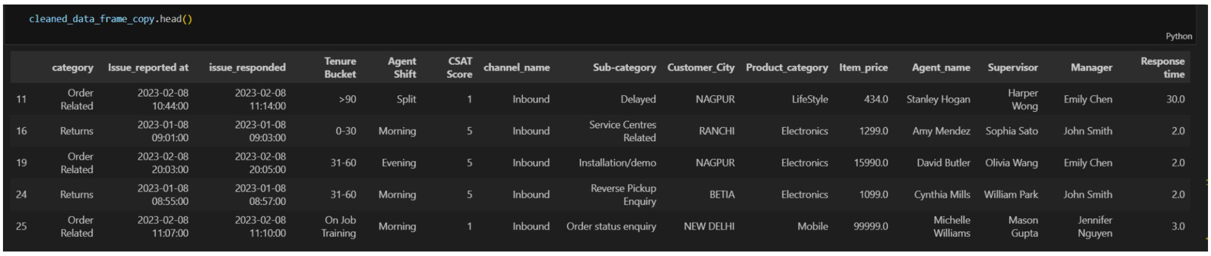
The connected handling time had a lot of missing values rendering it useless for our data exploration. Hence we decided to create a new variable called response time obtained by subtracting the issue reported and issue responded variables.
The cleaning part of the dataset has been completed.

(input

Statistical summary

Now we will be analysing the statistical summary of the cleaned dataset. We now check the count to display the unique values for each of the variables under the cleaned dataset.

While analysing the count we find that Customer city, Agent name and Supervisor has many different categorical values and hence would be too insignificant to consider each categorical value in evaluating CSAT value so we dropped it too.We will also be dropping the Item_price because there is no proper currency specified in the dataset for a good gauge of the value of the items.

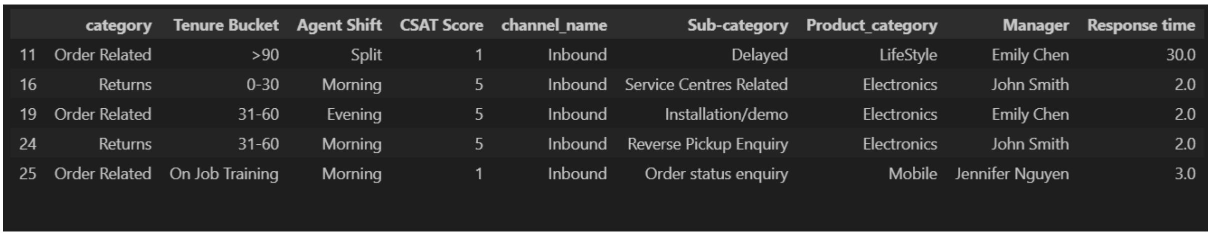


```
cleaned_data_frame_copy.head()
```

	category	Issue_reported at	issue_responded	Tenure Bucket	Agent Shift	CSAT Score	channel_name	Sub-category	Customer_City	Product_category	Item_price	Agent_name	Supervisor	Manager	Response time
11	Order Related	2023-02-08 10:44:00	2023-02-08 11:14:00	>90	Split	1	Inbound	Delayed	NAGPUR	LifeStyle	434.0	Stanley Hogan	Harper Wong	Emily Chen	30.0
16	Returns	2023-01-08 09:01:00	2023-01-08 09:03:00	0-30	Morning	5	Inbound	Service Centres Related	RANCHI	Electronics	1299.0	Amy Mendez	Sophia Sato	John Smith	2.0
19	Order Related	2023-02-08 20:03:00	2023-02-08 20:05:00	31-60	Evening	5	Inbound	Installation/demo	NAGPUR	Electronics	15990.0	David Butler	Olivia Wang	Emily Chen	2.0
24	Returns	2023-01-08 08:55:00	2023-01-08 08:57:00	31-60	Morning	5	Inbound	Reverse Pickup Enquiry	BETIA	Electronics	1099.0	Cynthia Mills	William Park	John Smith	2.0
25	Order Related	2023-02-08 11:07:00	2023-02-08 11:10:00	On Job Training	Morning	1	Inbound	Order status enquiry	NEW DELHI	Mobile	99999.0	Michelle Williams	Mason Gupta	Jennifer Nguyen	3.0

Data Visualization

We will now be comparing the cleaned dataset columns with the csat score. We will now be analysing the violin plot and boxplot for each variable.



	category	Tenure Bucket	Agent Shift	CSAT Score	channel_name	Sub-category	Product_category	Manager	Response time
11	Order Related	>90	Split	1	Inbound	Delayed	LifeStyle	Emily Chen	30.0
16	Returns	0-30	Morning	5	Inbound	Service Centres Related	Electronics	John Smith	2.0
19	Order Related	31-60	Evening	5	Inbound	Installation/demo	Electronics	Emily Chen	2.0
24	Returns	31-60	Morning	5	Inbound	Reverse Pickup Enquiry	Electronics	John Smith	2.0
25	Order Related	On Job Training	Morning	1	Inbound	Order status enquiry	Mobile	Jennifer Nguyen	3.0

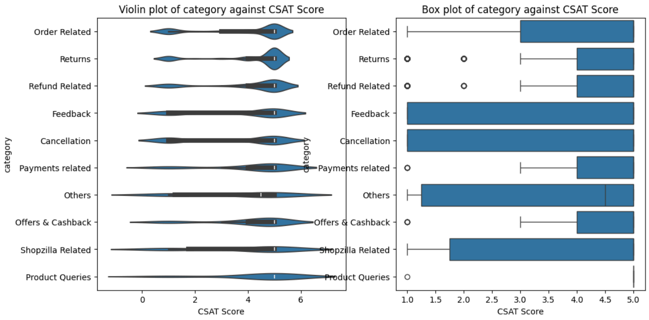


Figure 1

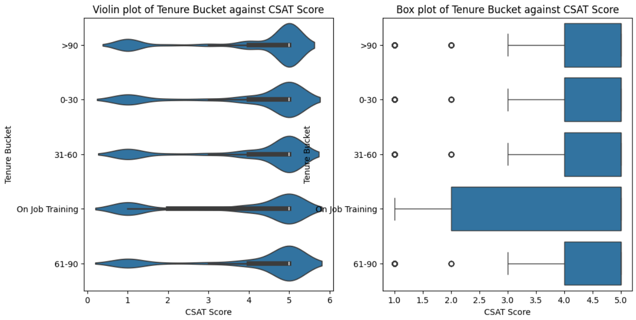


Figure 2

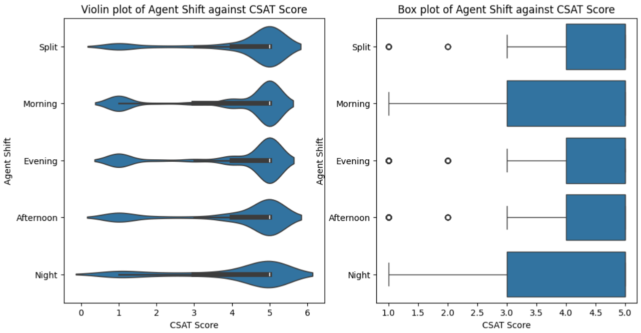


Figure 3

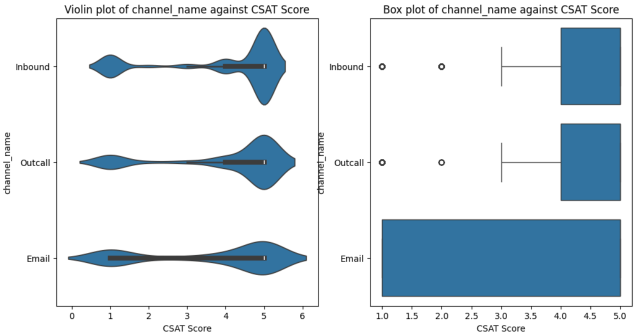


Figure 4

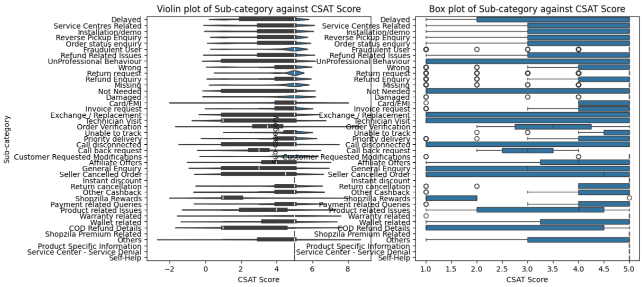


Figure 5

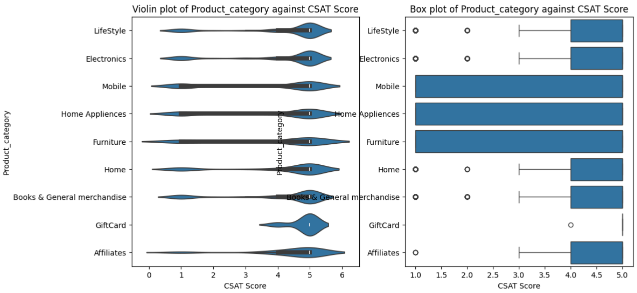


Figure 6

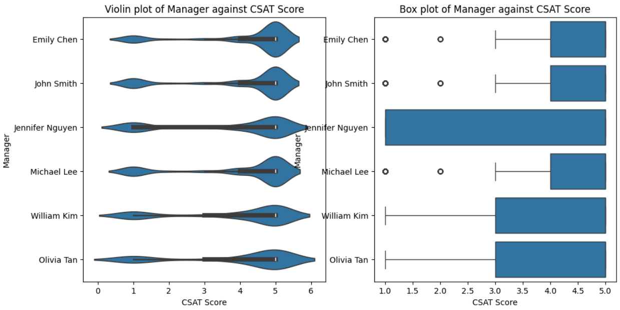


Figure 7

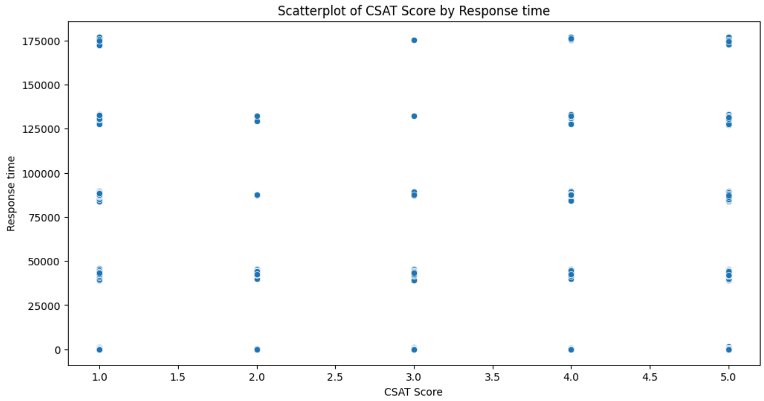


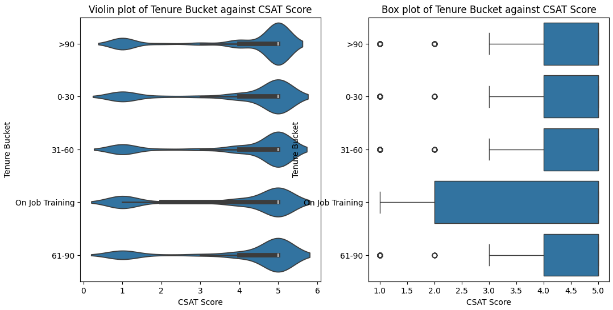
Figure 8

Category against csat scores

The category variable is more related to the category of the enquiry, having subdivisions like order related, returns, refund related, feedback and many more. First the category vs CSAT score shows us that its data points are too spread out and hence is not very good at predicting the csat score. Half of the category plot interquartile range spreads over from 1 to 5 csat score. Only Returns, refund related, payment related and offers and cashback are more uniform and show a high csat score of 4 to 5.

Tenure bucket vs csat scores

The tenure bucket indicates the duration of the employees at the company, varying from 30 days to more than 90 days or On job training. For the tenure bucket vs csat score the median csat score lies from 4 to 5 indicating a general positive satisfaction level across all shifts. Only OJT is slightly too spread out.

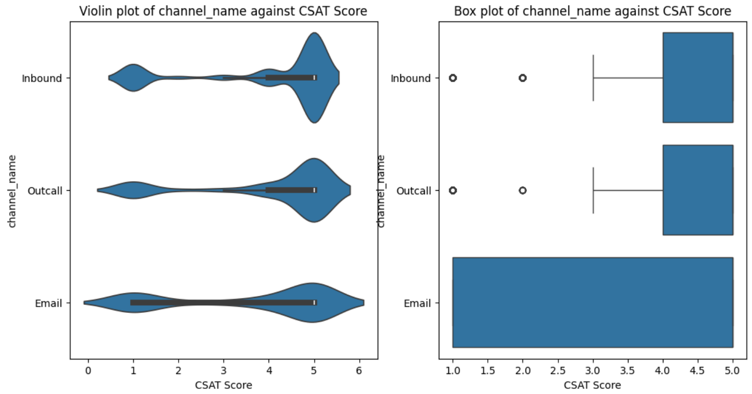


Agent shift vs CSAT scores

The agent shift is categorised into morning ,evening and afternoon shifts.We can infer from the plot that majority of the shift lies in the csat scores of 4 to 5.

Channel vs csat scores

Under the channel vs csat score plots the interquartile range for both the inbound and outcall is tightly clustered and between 4 and 5 whereas for emails there is a broader spread of satisfaction scores.

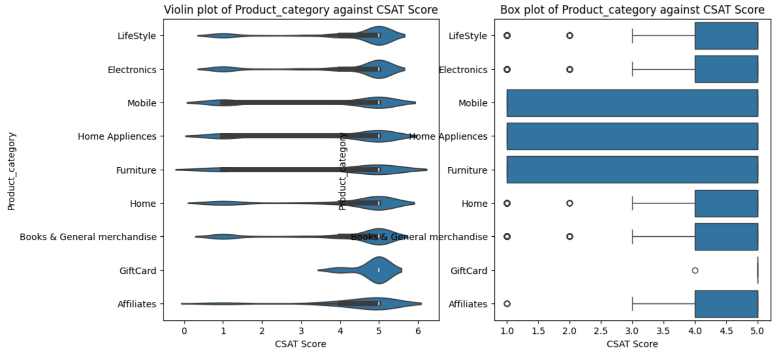


Subcategory vs csat scores

For the subcategory variable there are many subdivisions like delayed,services centres related, installation/demo, etc.The plotted figures shows us that each subdivisions spread across the CSAT scores non uniformly.There are quite a lot of subdivisions in the csat score from 1 to 5.While there some others in the 3 to 5 range.

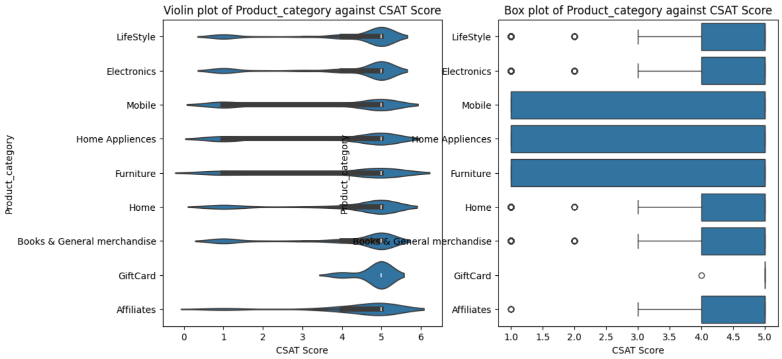
Product category vs csat

For the product category against csat scores plot majority has the interquartile range of 4 to 5. For mobile, home appliances and furniture, the data is too spread out to have a good trend.



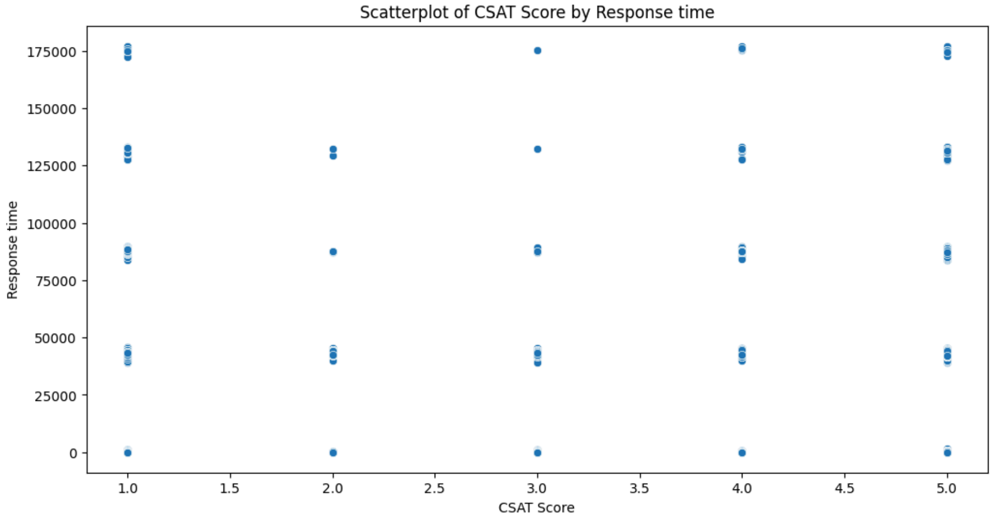
Manager vs csat

Manager against csat scores shows that half of the managers indicated high csat score which spans over from 4 to 5 whereas the rest is more spread between csat scores 3 and 5 and for Jennifer, the data is too spread out for any meaningful trend.



Response time vs csat

The response time scatter plot is spread out all over the plot, indicating weak relation between the 2 variables.



<u>Principal Component analysis</u>

We will be using **Principal Component Analysis** to simplify the multiple data we have to train the machine for prediction of CSAT score.

Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify complex datasets by transforming them into a lower-dimensional space while preserving most of the essential information.

Data Transformation: PCA transforms the original feature space into a new orthogonal (uncorrelated) feature space.

Dimensionality Reduction: It identifies the principal components (PCs), which are linear combinations of the original features. These PCs capture the maximum variance in the data. Ordering of Components: The first principal component (PC1) explains the most variance, followed by PC2, PC3, and so on. Each subsequent component explains less variance than the previous one.

Variance Retention: PCA retains as much variance as possible while reducing the dimensionality of the data. Typically, only the top principal components that explain most of the variance are kept, while the rest are discarded.
Orthogonality: The principal components are orthogonal to each other, meaning they are uncorrelated. This property simplifies the interpretation of the transformed data.

For our dataset, the remaining cleaned data is simplified into 2 Principal Components.

Principal Component 1 (PC1):
This component captures the largest amount of variance in the data. It is primarily influenced by variables like channel_name, category, Response time, and Tenure Bucket, as indicated by their relatively high loadings. PC1 represents the main patterns or trends in the data that are shared among these variables.

Principal Component 2 (PC2):
PC2 captures the second-largest amount of variance in the data, orthogonal to PC1. It is influenced by variables such as Tenure Bucket, channel_name, and Response time, although some of these variables have negative loadings, suggesting an inverse relationship. PC2 represents additional patterns or trends in the data that are not captured by PC1.

Categorical to numerical (Data Optimisation)
Because most data comparison and machine learning algorithms works with numerical values, we had to use the one hot encoder function to transform the categorical data into numerical, binary representation to be read by the algorithm

Explained Variance of PCA Variables
The explained variance ratio of a principal component is the proportion of the dataset's variance explained by that component alone. It indicates the amount of information (variance) retained by each principal component.

The explained variance ratio of [0.0019574, 0.0016786] means that the first principal component explains approximately 0.19574% of the variance in the original data, while the second principal component explains approximately 0.16786% of the variance.

When converted to percentages, these values are approximately 0.19574% and 0.16786%, respectively.

These percentages are quite low, indicating that the two principal components capture only a small amount of the total variance in the data. This suggests that the original variables may not be well represented by the principal components, and the model may not be effective in reducing the dimensionality of the data.

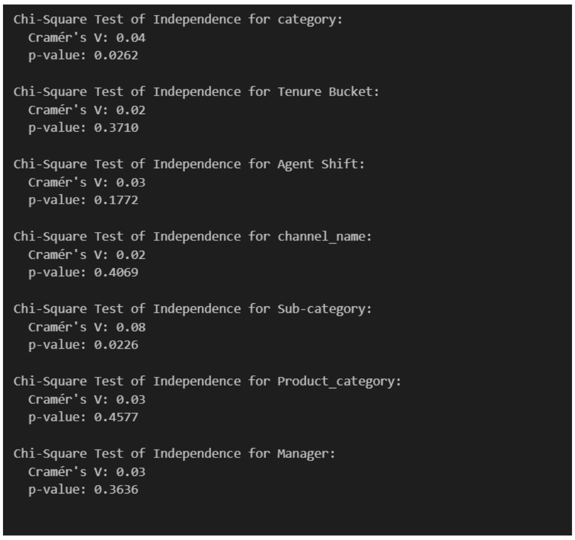
PCA refining
Due to the poor explained variance ratio, we have decided to eliminate the lower bearing variables and keep only the variables that have significant association with CSAT variable

To do this, we evaluated the loading percentage of each variable into each PCA variable and we looked at the Chi Square Test of independence and correlation values of the variables.

Chi Square Test of Independence
Why did we use it? Correlation is more suited for numerical values and we’re comparing 2 categorical variables, the initial variables and CSAT score.
This test assesses whether there is a significant association between two categorical variables.

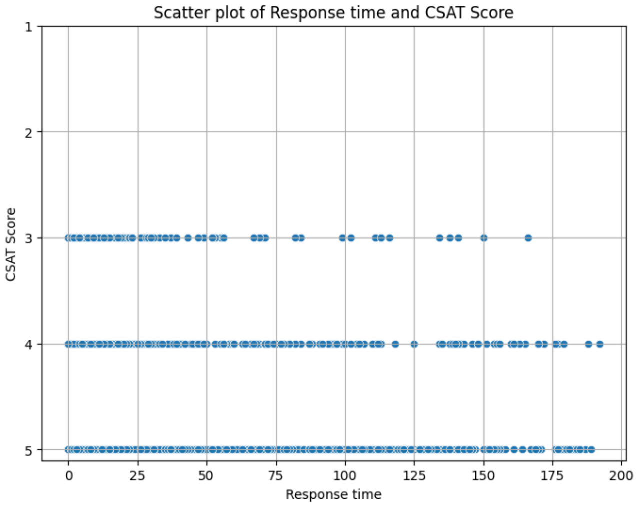
Cramér's V: Cramér's V is a measure of association between two categorical variables. It ranges from 0 to 1, where 0 indicates no association and 1 indicates a perfect association.

The P Value tells us the likelihood of observing our data if there is no real effect or relationship in the population. In other words, it quantifies the evidence against the null hypothesis.



```
Chi-Square Test of Independence for category:
  Cramér's V: 0.04
  p-value: 0.0262

Chi-Square Test of Independence for Tenure Bucket:
  Cramér's V: 0.02
  p-value: 0.3710

Chi-Square Test of Independence for Agent Shift:
  Cramér's V: 0.03
  p-value: 0.1772

Chi-Square Test of Independence for channel_name:
  Cramér's V: 0.02
  p-value: 0.4069

Chi-Square Test of Independence for Sub-category:
  Cramér's V: 0.08
  p-value: 0.0226

Chi-Square Test of Independence for Product_category:
  Cramér's V: 0.03
  p-value: 0.4577

Chi-Square Test of Independence for Manager:
  Cramér's V: 0.03
  p-value: 0.3636
```

We used ***correlation to find the relationship between Response time and CSAT*** value as it is a numerical variable.



OPTIMISATION RESULTS

Variables like "category," "Sub-category,"have relatively low loadings across both principal components and may have weak associations with "CSAT Score" based on the Chi-square test, hence they are removed from consideration in PCA.

PCA RESULTS

There’s a slight increase in the explained variance percentage from initially 0.19574% in PC1 to 1.2986% and from 0.16786% in PC2 to 1.1827%

Final Prediction results

Now, we used the PCA Variables to train our machine with a Decision tree algorithm, to predict CSAT Value. We also used the initial cleaned data variables, from the columns of the dataset to train a separate decision tree algorithm to note the differences in prediction of the score.

WHAT IS DECISON TREE

The Decision Tree algorithm is a supervised machine learning algorithm used for both classification and regression tasks.

IT builds a tree-like structure to predict the target variable by recursively splitting the data based on features. Each leaf node represents a class label or numerical value. It's simple to understand and interpret but can overfit with deep trees.

We used F1 SCORE AND ACCURACY to evaluate the effectiveness. What are they?

Accuracy: Accuracy measures the proportion of correct predictions out of the total number of predictions made by the model. It is calculated as the number of correct predictions divided by the total number of predictions. Accuracy gives an overall indication of how often the model correctly predicts the outcome across all classes.

F1 Score: F1 score is the harmonic mean of precision and recall. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, while recall measures the proportion of true positive predictions out of all actual positive instances in the dataset. F1 score balances both precision and recall and is particularly useful when dealing with imbalanced datasets where one class is much more frequent than the other.

The PCA trained prediction model did slightly better than if we were to just use a multivariate prediction model of the decision tree.

The Accuracy score and the F1 score of PCA was higher at
F1 Score: 0.7129765164158266
Accuracy: 0.735966735966736

While these were for non PCA
F1 Score: 0.7014626783699713
Accuracy: 0.7182952182952183

The PCA Decision tree model has a relatively good prediction of the CSAT SCORE at about 70% accuracy and a good F1 score of also about 0.71, meaning that there's a good balance performance between precision and recall in a classification task

Due to the multiple data variables and hence multiple variances in the data from each data and their corresponding categorical values and numerical values, the final decision tree has many nodes that each leads to a predictive value for the categorical data CSAT score.

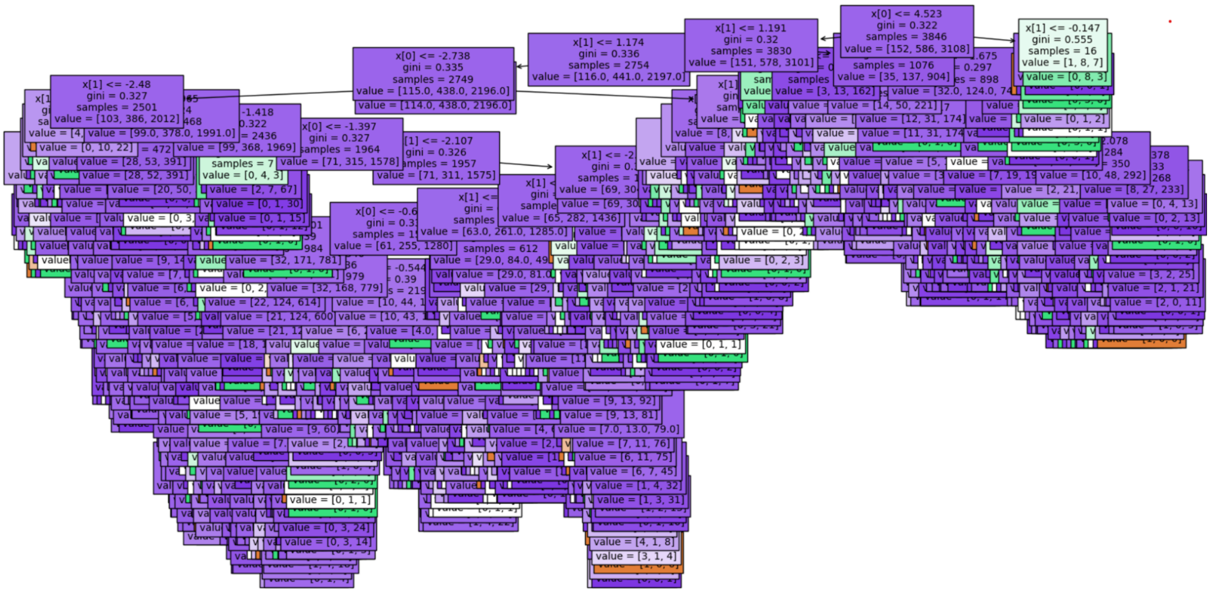
Why are PCA Decision Tree predictions better than the multivariate one?

Decision tree learning with PCA variables may sometimes better compared to using the original multivariate variables for several reasons:

Dimensionality reduction: PCA reduces the dimensionality of the feature space by transforming the original variables into a smaller set of principal components. This can lead to simpler decision trees with fewer nodes and splits, making them easier to interpret and less prone to overfitting.

Noise reduction: PCA tends to remove noise and redundant information from the dataset, focusing on the principal components that capture the most significant variability in the data. This can result in decision trees that are more robust and generalise better to unseen data.

Improved interpretability: PCA variables often represent combinations of the original variables that capture underlying patterns or structures in the data. Decision trees built on these variables may uncover more interpretable relationships between features and the target variable.



Conclusion of project:
From Principal Component Analysis, the 2 variables that are most significant in capturing and influencing final CSAT score are the Response Time, with a negative correlation, and the Channel Name (inbound, outcall, email etc), with a positive correlation. The response time is a clear and logical component that affects CSAT score as how long the interaction takes directly translates to the efficiency of the problem resolution. Quick responses boost satisfaction levels, while delays may lead to dissatisfaction.

The Channel Name, as depicted by the exploratory analysis, calls, inbound and outbound do better in CSAT score at 4-5 which is high, and emails are more spread out. The channel name component and its influence, could be a result of human psychology where telephone calls are more personal whereas while emails are instantaneous, telephoning someone means taking time out of your day to stop and make the call. This shows more care, demonstrates more attention and better can demonstrate better customer service.
Secondly, Phone calls are often faster, Messages can be conveyed more quickly over calls than exchanging them via email, making the interactions more efficient.

The result of our project indicates that these 2 components are something that the company Shopzilla should consider more on, to reduce response time and focused channel name on calls instead of emails, especially for conflict resolution.
Mainly, fast response time increases customer satisfaction by providing quick solutions to problems and reducing issue resolution time. A fast response to customer queries builds trust and loyalty, demonstrating a commitment to excellent customer service.