DSAI
Mini Project

E Commerce Customer Service Satisfaction

GROUP 4

YUN TAT RUGMA SAAD

Target Company

shopzilla®

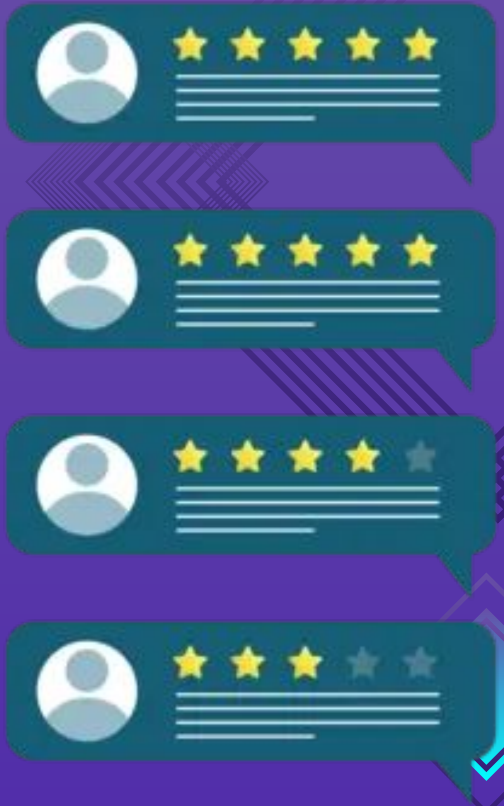
Shopzilla, now known as Connexity, is a company focused on online retail and ecommerce.

It provides a platform for online shopping, connecting shoppers with retailers and offering a wide range of products

The company primarily serves the ecommerce industry. It was founded in 1996 and is based in Los Angeles, California.

About Our Dataset

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 interaction with the client affect the ultimate satisfaction rating of the customer in ecommerce shopping?

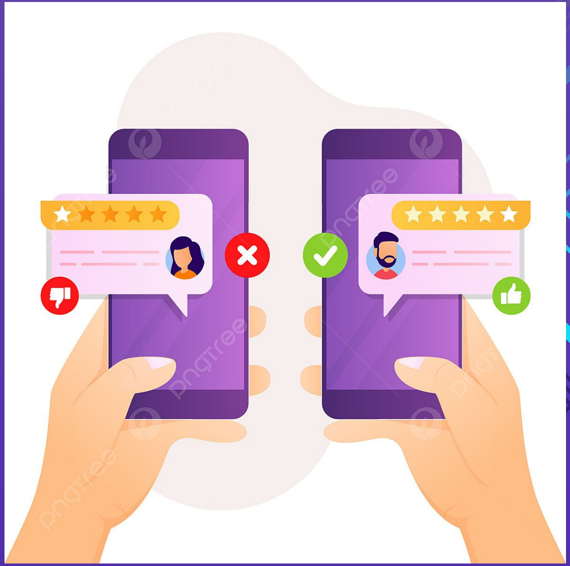
Significance of the problem

Customer Reviews are crucial to the success of a business.

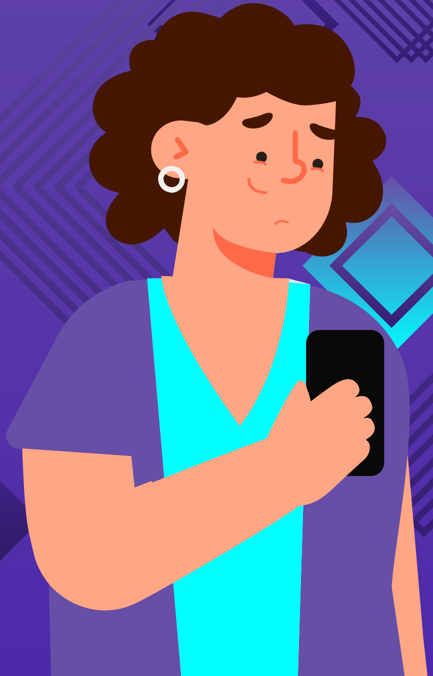
- Psychologically, we place significant value on others' opinions and behaviors.
- Decision-making is often influenced by the choices of others.

Customer reviews act as personal testimonials.

- They engage readers in a dialogue with the reviewer and indirectly with the brand.
- Reviews offer insights into product or service experiences.
- They influence decision-making through affective forecasting.



Data Cleaning & Exploration



Dataset Columns

- Category
- Issue_reported at
- Issue_responded
- Tenure Bucket
- Agent Shift
- Channel_name
- Sub-category
- Customer_City
- Product_category
- Item_price
- Agent_name
- Supervisor
- Manager

Checking for Missing Values

We chose 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.

```
cleaned_data_frame = df[important_columns]
missing_values = cleaned_data_frame.isnull().sum()
print(missing_values)
```

```
category                0
Issue_reported at       0
issue_responded         0
Tenure Bucket           0
Agent Shift             0
CSAT Score              0
[illegible]
Sub-category            0
Customer_City       68828
Product_category    68711
Item_price          68701
Agent_name              0
Supervisor              0
Manager                 0
dtype: int64
```

Dataset Columns

- Category
- Issue_reported at
- Issue_responded
- Tenure Bucket
- Agent Shift
- Channel_name
- Sub-category
- ~~Customer_City~~
- ~~Product_category~~
- ~~Item_price~~
- Agent_name
- Supervisor
- Manager

New Variable

Time of issue responded- Time of Issue reported

= RESPONSE TIME

Statistical Summary

1) Numeric Variables like Response time - **mean, standard deviation** etc.

2) Categorical Variables - check the **count to display the unique values** for each of the categorical variables under the cleaned dataset.

Dropping more variables

- Customer city, Agent name and Supervisor (due to too many categorical values & would be too insignificant to include them)
- Item_price (no proper currency specified)

```
category
Returns              3971
Order Related        2471
Refund Related        518
Cancellation          346
Feedback              218
Offers & Cashback      35
Payments related       31
Others                 14
Shopzilla Related      14
Product Queries         6
Name: count, dtype: int64
```

```
Tenure Bucket
>90                 3018
31-60               1351
On Job Training     1324
0-30                1171
61-90                760
Name: count, dtype: int64
```

```
Supervisor
Elijah Yamaguchi    412
Carter Park         396
Noah Patel          386
Nathan Patel        337
Emma Park           310
Zoe Yamamoto        306
William Park        295
Madison Kim         294
Mia Patel           294
Evelyn Kimura       285
Aiden Patel         276
Scarlett Chen       276
Logan Lee           263
Jackson Park        236
Brayden Wong        221
Lily Chen           210
Emily Yamashita     193
Ava Wong            188
Olivia Wang         184
Mason Gupta         169
Landon Tanaka       165
Amelia Tanaka       159
Sophia Sato         146
Olivia Suzuki       145
...
Sophia Chen          10
Name: count, dtype: int64
```

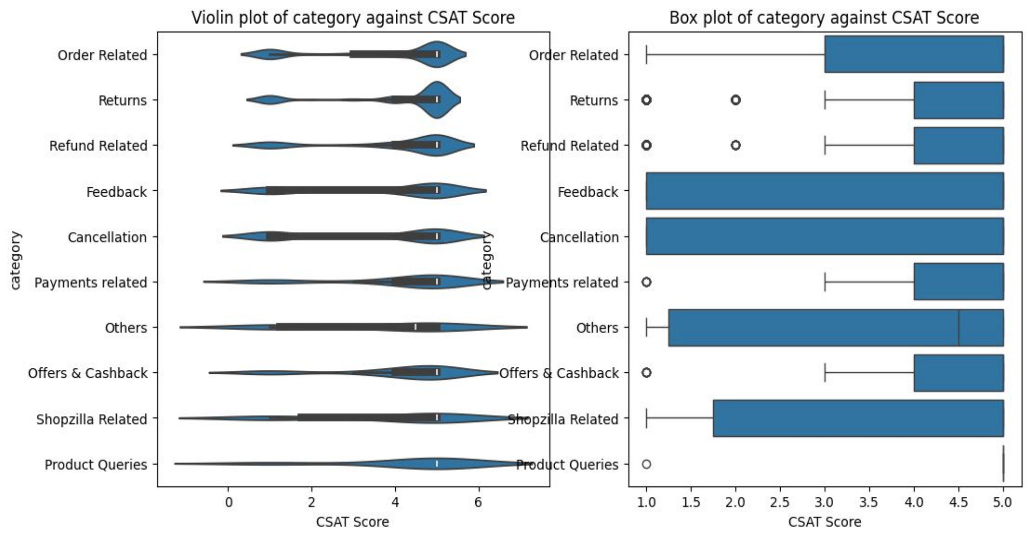
```
count      7624.000000
mean      10697.920121
std       27111.619772
min           0.000000
25%           2.000000
50%           8.000000
75%         140.250000
max      177097.000000
Name: Response time, dtype: float64
```

Dataset Columns

- Category
- ~~Issue_reported at~~
- ~~Issue_responded~~
- Tenure Bucket
- Agent Shift
- Channel_name
- Response time ***
- Sub-category
- ~~Customer_City~~
- ~~Product_category~~
- ~~Item_price~~
- ~~Agent_name~~
- ~~Supervisor~~
- Manager

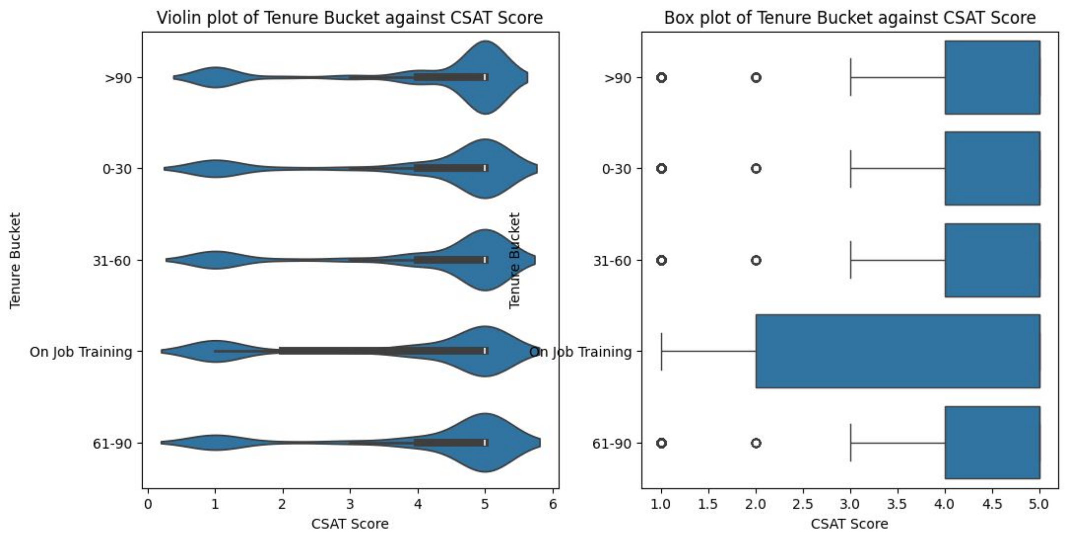
Data Visualization

Category against CSAT Scores



- Half of the category plot interquartile range spreads over from 1 to 5 CSAT Score.
- Only returns, refund related, payment related, offers and cashback are more uniform and show a high csat score of 4 to 5.

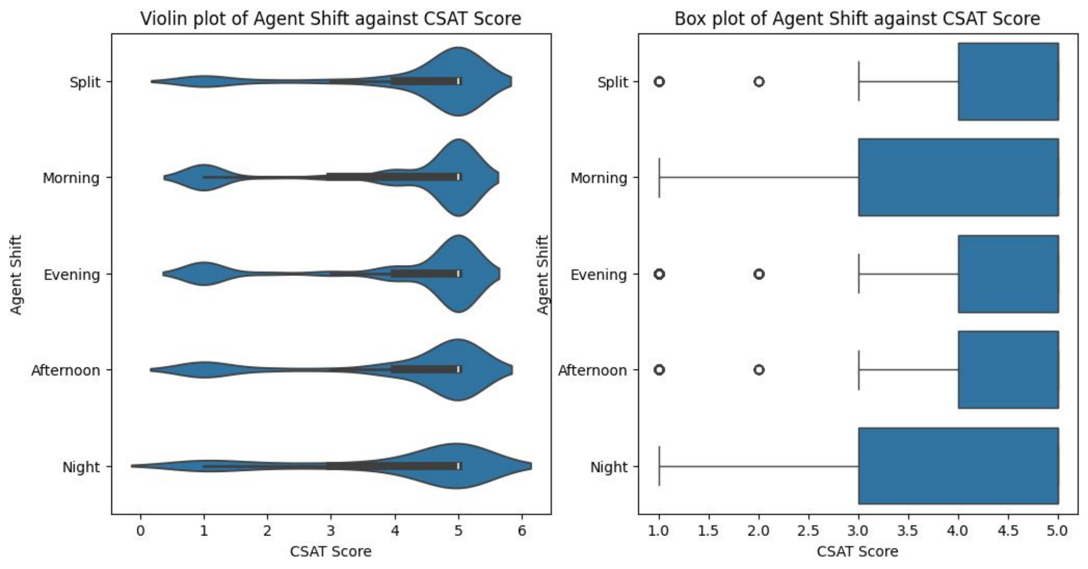
Tenure Bucket vs CSAT



- The tenure bucket indicates the duration of the employees at the company, varying from 30 days to more than 90 days or on-the-job training.

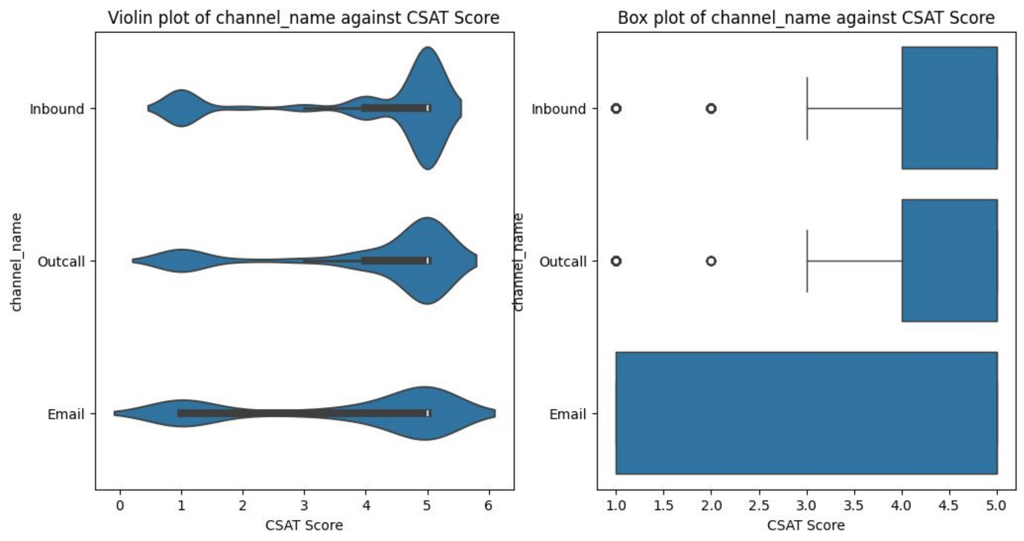
- For the tenure bucket vs CSAT scores, the median CSAT score lies from 4 to 5 indicating a general positive satisfaction level across all shifts.

Agent shift vs CSAT



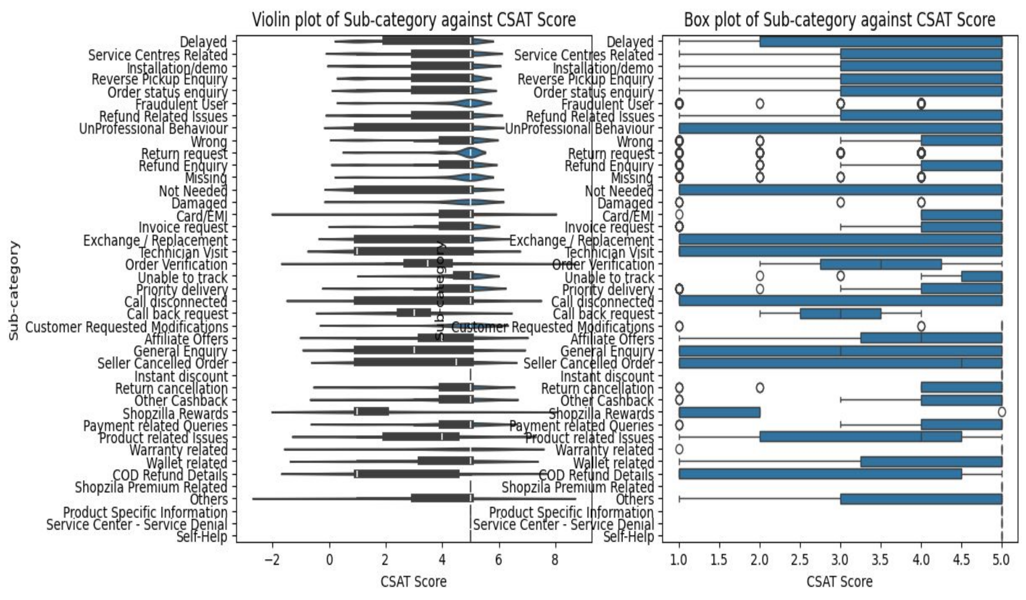
- The agent shift refers to the time period which the agent is working.
- We can infer from the plot that majority of the shift lies in the csat scores of 4 to 5.

Channel vs CSAT



- 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.

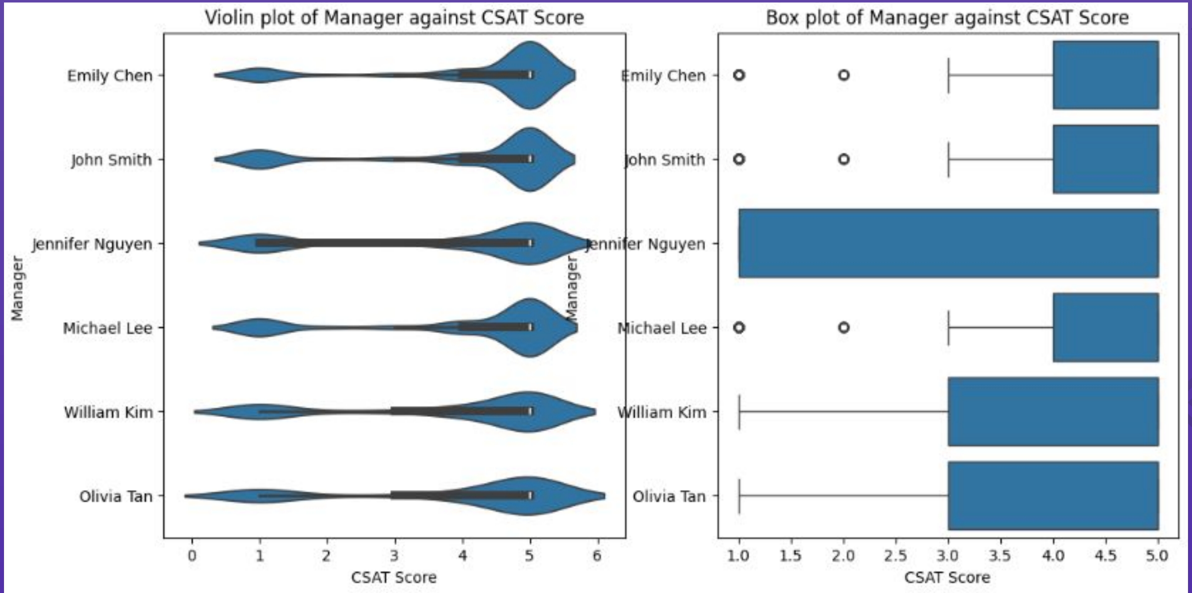
Subcategory vs CSAT



- For the subcategory variable there are many subdivisions like delayed, services centres related, installation/demo, etc.

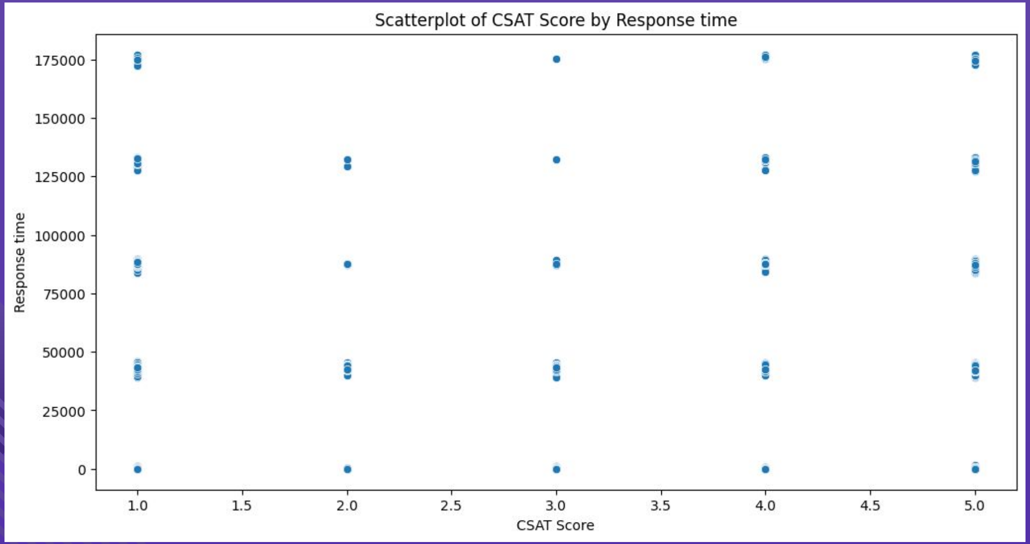
- The plotted figures shows us that each subdivisions spread across the CSAT scores non uniformly. There are quite a lot of subdivisions in the csat score from 1 to 5. While there are some others in the 3 to 5 range.

Manager vs CSAT



- 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 .
- The data is too spread out for any meaningful trend.

Response Time vs CSAT Scores



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

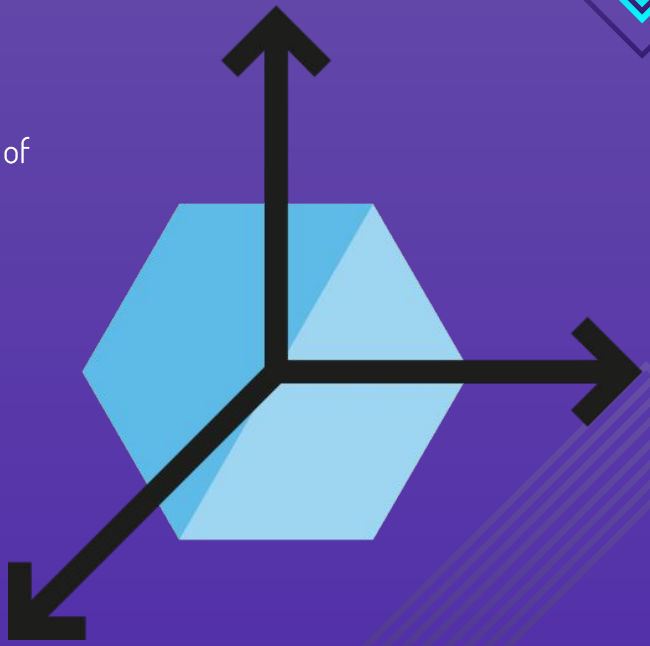
Principal Component Analysis

dimensionality reduction technique used to simplify complex datasets by transforming them into a lower-dimensional space while preserving most of the essential information.

1) Data Transformation

2) Dimensionality Reduction

3) Variance Retention



Principal Component Variables

PC1	PC2

PC1

- Largest Variance
- Main patterns or trends

PC2

- Second largest Variance
- Orthogonal to PC1
- Additional patterns or trends in data not captured by PC1

ONE HOT ENCODER OPTIMISATION

Categorical -> Numerical

```python
# One-hot encode categorical variables
encoder = OneHotEncoder()
X_encoded = encoder.fit_transform(X)
```

Explained Variance of PCA Variables

0.19%

PC1

0.16%

PC2

Chi Square Test of Independence

Why use it?

- This test assesses whether there is a significant association between two categorical variables.

Components

- Cramér's V
- P Value

Optimisation Results

```
Principal Component 1:
Variable 1: category, Loading: 0.05593541602491599
Variable 3: Agent Shift, Loading: 0.026606387301456154
Variable 5: Sub-category, Loading: 0.019780536082372485
Variable 7: Manager, Loading: 0.056432505412546775
Principal Component 2:
Variable 1: category, Loading: 0.038793199256909494
Variable 2: Tenure Bucket, Loading: -0.15474403881870283
Variable 5: Sub-category, Loading: -0.006802693793666027
Variable 7: Manager, Loading: -0.09869730175459725
Variable 8: Response time, Loading: 0.14686548748630412
```

```
Correlation between Response time and CSAT Score: -0.0660021752867596
```

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

  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

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

  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
```

Explained Variance of PCA Variables

1.29%

PC1

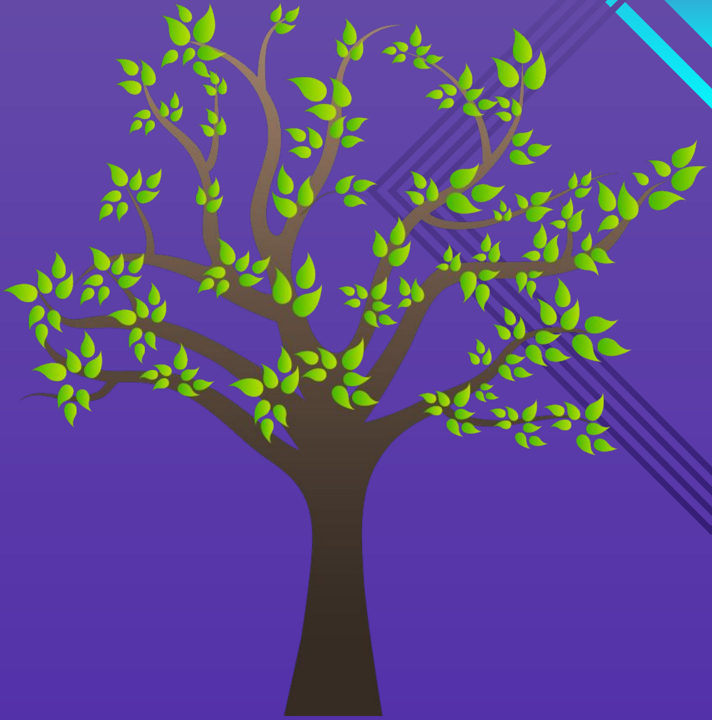
1.18%

PC2

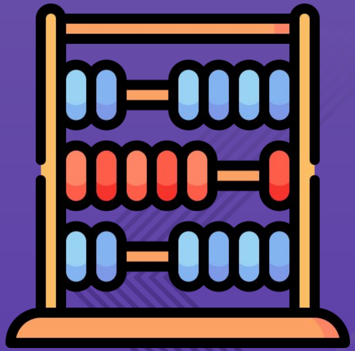
Dataset Columns

- ~~Category~~
- ~~Issue_reported at~~
- ~~Issue_responded~~
- Tenure Bucket
- Agent Shift
- Channel_name
- Response time ***
- ~~Sub-category~~
- ~~Customer_City~~
- ~~Product_category~~
- ~~Item_price~~
- ~~Agent_name~~
- ~~Supervisor~~
- Manager

DECISION TREE ALGORITHM



F1 Score & Accuracy



Accuracy:

- Measures correct predictions out of total predictions.
- Provides an overall indication of model performance across all classes.

F1 Score:

- Harmonic mean of precision and recall.
- Balances precision (true positive predictions out of all positive predictions) and recall (true positive predictions out of all actual positive instances).

F1 Score & Accuracy

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

PCA:
F1 Score: 0.7129765164158266
Accuracy: 0.735966735966736

Non-PCA:
F1 Score: 0.7014626783699713
Accuracy: 0.7182952182952183

The PCA Decision tree model has a good prediction of the CSAT SCORE at about 70% accuracy and a good F1 score of also about 0.71,

- good balance performance between precision & recall in a classification task

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

Dimensionality reduction:

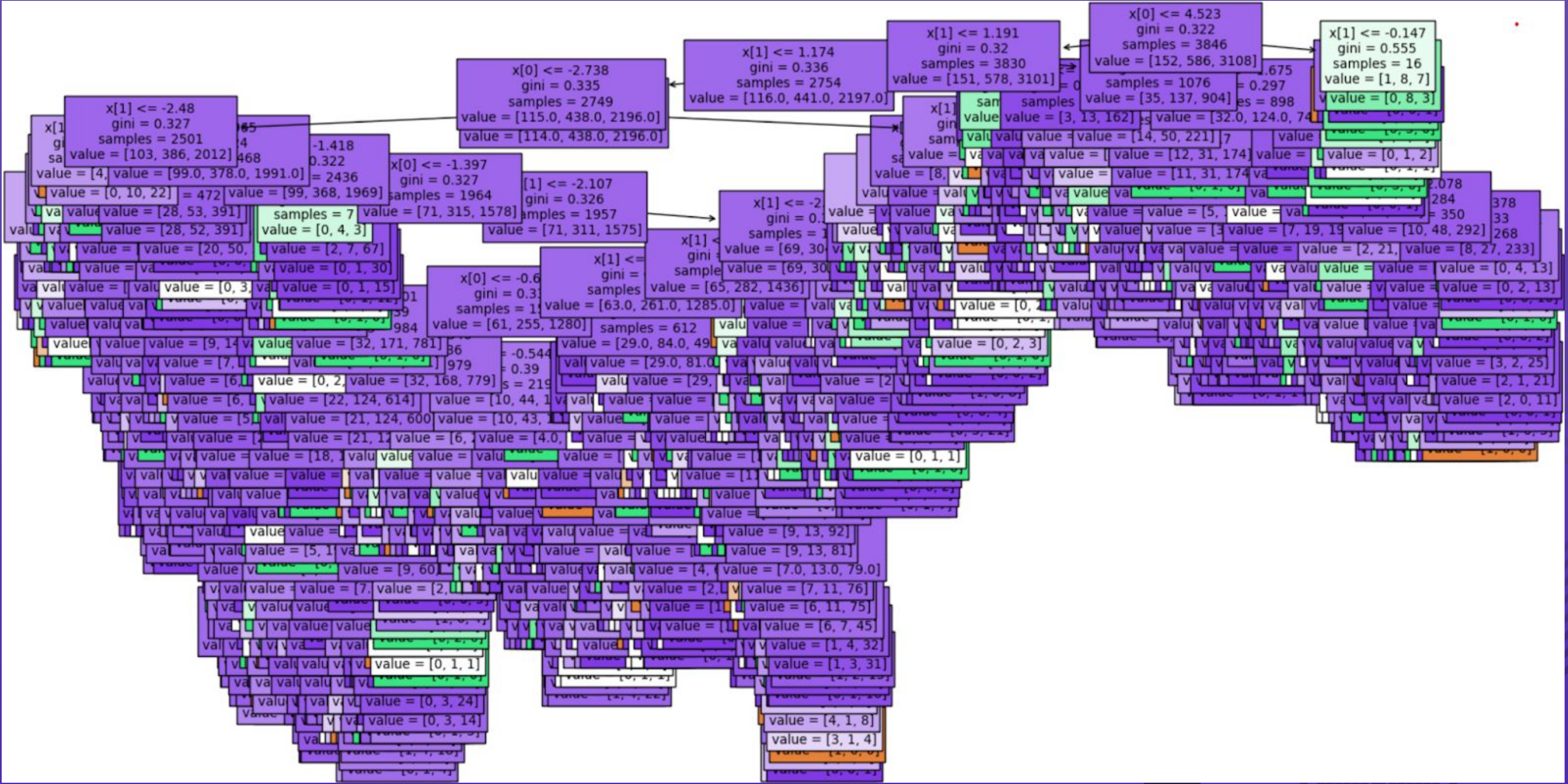
- PCA reduces feature space
- Transforms variables into fewer principal components
- Results in simpler and more predictable decision trees.

Noise reduction:

- PCA removes noise and redundancy
- Focuses on capturing significant data variability
- Results in robust and generalizable decision trees

Improved interpretability:

- PCA variables combines original variables
- Helps decision trees reveal interpretable feature target relationships.


x[0] <= 4.523
gini = 0.322
samples = 3846
value = [152, 586, 3108]
x[1] <= 1.191
gini = 0.32
samples = 3830
value = [151, 578, 3101]
x[1] <= -0.147
gini = 0.555
samples = 16
value = [1, 8, 7]
x[1] <= 1.174
gini = 0.336
samples = 2754
value = [116.0, 441.0, 2197.0]
x[0] <= -2.738
gini = 0.335
samples = 2749
value = [115.0, 438.0, 2196.0]
value = [114.0, 438.0, 2196.0]
x[1] <= -2.48
gini = 0.327
samples = 2501
value = [103, 386, 2012]
x[0] <= -1.397
gini = 0.327
samples = 1964
value = [71, 315, 1578]
value = [71, 311, 1575]
value = [65, 282, 1436]
value = [63.0, 261.0, 1285.0]
value = [61, 255, 1280]

CONCLUSION

2 variables that are most significant in capturing and influencing final **CSAT scoreS** are:

- **Response Time** (negative correlation)
- **Channel Name** (inbound, outcall, email etc), with a positive correlation.

Response Time

- Clear and logical
- Duration proportional to efficiency
- Quick responses boost csat

Channel Name

- Inbound and Outbound calls have higher csat
- Calls are more personal and faster
- Emails have more spread out csat
- Emails are more formal and slower

Thank You!