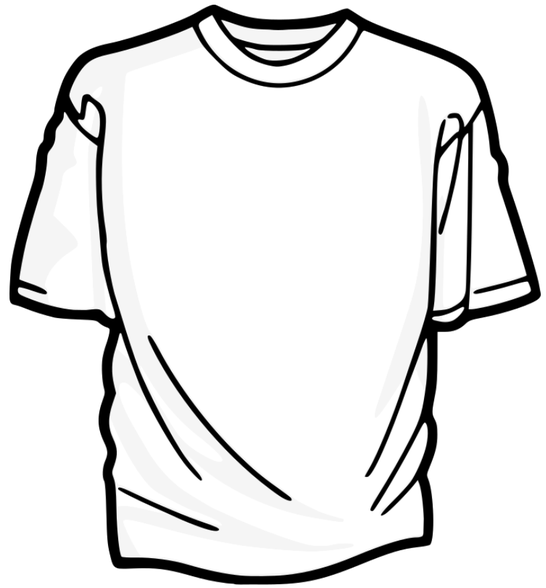
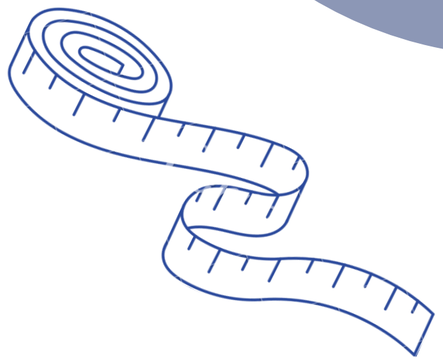


PRODUCT SIZE RECOMMENDATION

ADITI VASA - AU2040122
SHREY SOMANI - AU2040002
VANDAN SHAH - AU2040196
RONIT SHAH - AU2040048

INTRODUCTION & PROBLEM STATEMENT

Problem Statement:
Tackle the problem faced by customers regarding cloth sizes while shopping



- **Main goal:** Help people find their suitable size that fits them
- Model that can predict the size of clothing
- Train model based on collected dataset of body measurements and corresponding clothing sizes
- Various algorithms can be used like logistic regression, random forest, etc. Mainly classifier algorithms
- Use the training set to make predictions on new data
- Evaluate performance on testing set

EXISTING BODY OF WORK

- *Relatively new concept so only a few researches available*
- *One of the method uses common classification problem*
- *Other prediction models include KNN, Random Forest*
- *A recent method learns the latent properties of customers and products using skip-gram models.*
- *A research also suggests using logistic regression to give ordinal categories for improved model fitting*

OUR APPROACH

1

DATA CLEANING

Remove null values or missing values and scale the features using data imputation

2

EXPLORATORY DATA ANALYSIS

Visually understanding the data using different plots. Find correlation between various variables

3

ENCODING

Dealing with categorical data and applying ordinal encoding on ordinal data like 'fit' and one hot encoding on nominal data like 'body type'

4

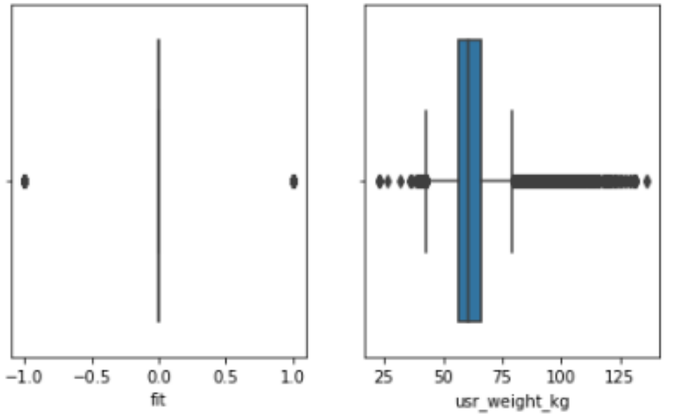
LOGISTIC REGRESSION

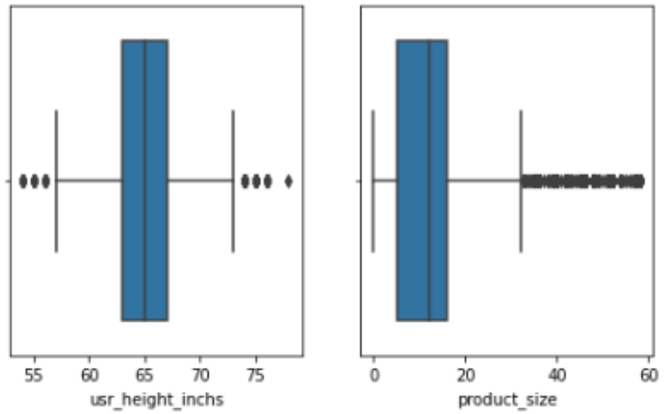
Capture fit semantics and use logistic regression algorith to predict the accuracy of the model

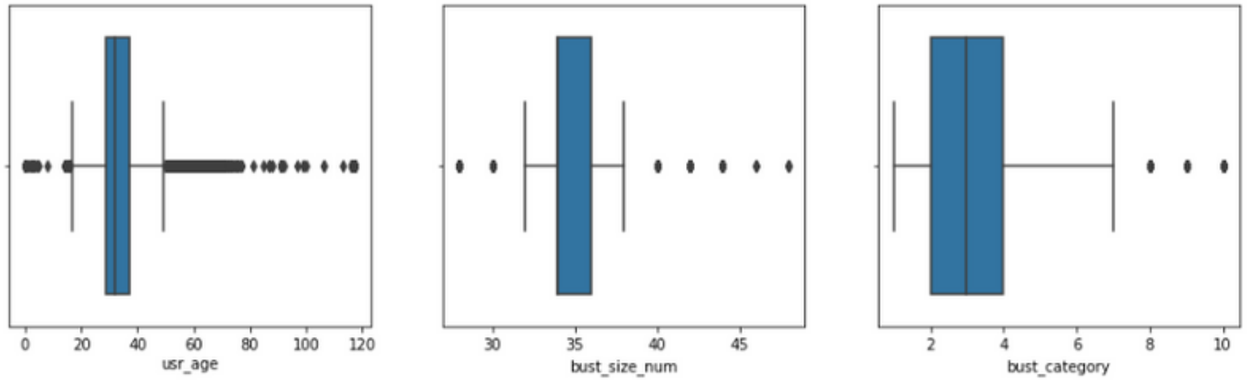
INITIAL RESULTS

```
def draw_boxplots(cols, data, per_line=4):
    n = len(cols)
    per_line = 4
    for i in range(0, n, per_line):
        n_plots = per_line if n - i >= per_line else n % per_line
        fig, axes = plt.subplots(1, n_plots)
        plt.subplots_adjust(wspace=0.2)
        fig.set_figwidth(15)
        fig.set_figheight(4)
        for j in range(n_plots):
            sns.boxplot(data[cols[i + j]], ax=axes[j])
        plt.show()

draw_boxplots(numeric_features, newdf)
```





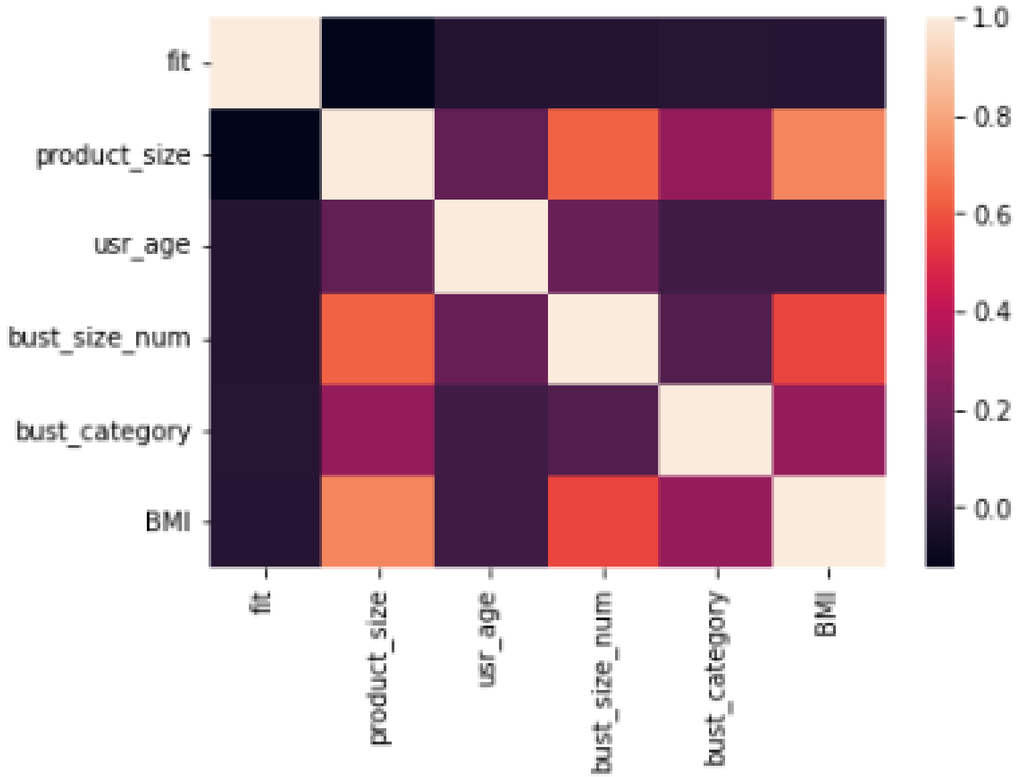


INITIAL RESULTS

```
cleaned_df_scaled = newdf[numeric_features].copy()
cleaned_df_scaled = pd.DataFrame(scale(cleaned_df_scaled), columns=numeric_features)

corr_matrix = cleaned_df_scaled.corr()
sns.heatmap(corr_matrix)
corr_matrix
```

	fit	product_size	usr_age	bust_size_num	bust_category	BMI
fit	1.000000	-0.124194	-0.016582	-0.015816	-0.008531	-0.010379
product_size	-0.124194	1.000000	0.161174	0.627775	0.297101	0.716304
usr_age	-0.016582	0.161174	1.000000	0.177865	0.067199	0.068538
bust_size_num	-0.015816	0.627775	0.177865	1.000000	0.113405	0.563523
bust_category	-0.008531	0.297101	0.067199	0.113405	1.000000	0.295737
BMI	-0.010379	0.716304	0.068538	0.563523	0.295737	1.000000



INITIAL RESULTS

```
# split X and y into training and testing sets
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=16)

logisticRegr = LogisticRegression()
logisticRegr.fit(X_train, y_train)
```

```
predictions = logisticRegr.predict(X_test)
score = logisticRegr.score(X_test, y_test)
print(score*100,"%")
```

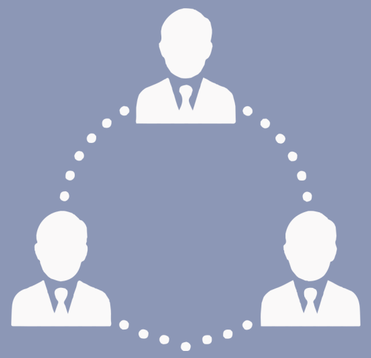
```
68.10769814182784 %
```

Accuracy of the model after applying Logistic Regression came out to be 68%. We aim to improve the accuracy for predicting future dataset

ROLE & FUTURE WORK

CONTRIBUTIONS

Aditi Vasa - Encoding and Report
Shrey Somani - EDA and Report
Vandan Shah - Regression and Presentation
Ronit Shah - Data Cleaning and Presentation



- Further we will implement more algorithms and compare which algorithm better suits the model and predicts the future data set accurately.
- We plan to do further EDA
- Other algorithms include QDA, Random Forest, Ridge Regresion, SVM, KNN etc

REFERENCES

CLOTHING FIT DATASET FOR SIZE RECOMMENDATION. (2018, AUGUST 21). KAGGLE. HTTPS://WWW.KAGGLE.COM/DATASETS/RMISRA/CLOTHING-FIT-DATASET-FOR-SIZE-RECOMMENDATION

MISRA, RISHABH, MENGTING WAN, AND JULIAN MCAULEY. "DECOMPOSING FIT SEMANTICS FOR PRODUCT SIZE RECOMMENDATION IN METRIC SPACES." IN PROCEEDINGS OF THE 12TH ACM CONFERENCE ON RECOMMENDER SYSTEMS, PP. 422-426. 2018.

MISRA, RISHABH AND JIGYASA GROVER. "SCULPTING DATA FOR ML: THE FIRST ACT OF MACHINE LEARNING." ISBN 9798585463570 (2021).

VIVEK SEMBIUM, RAJEEV RASTOGI, ATUL SAROOP, AND SRUJANA MERUGU. 2017. RECOMMENDING PRODUCT SIZES TO CUSTOMERS. IN RECSYS.

CAMPUSX. (2021, APRIL 13). ONE HOT ENCODING | HANDLING CATEGORICAL DATA | DAY 27 | 100 DAYS OF MACHINE LEARNING [VIDEO]. YOUTUBE. HTTPS://WWW.YOUTUBE.COM/WATCH?V=U5OCV3JKWKA