PREDICTIVE ANALYSIS OF AUDI CAR PRICES USING MACHINE LEARNING

Name: Sahil Kumar

Student ID: GH1036801

Module: Business Project in Big Data and AI

Module Code: M516

Github Repository Link: https://github.com/sahilkumar2306/M516-Audi-Car-Price-Prediction

Video Link: https://youtu.be/gyj_HVmJc6k

Table of Contents

1. Introduction

It is an initiative that explores the pricing mechanism of Audi cars found in the second-hand car market through the use of data science and machine learning. Knowledge of price determinants of vehicles is of significant commercial importance to the automotive dealers, second-hand car dealers, and consumers interested in the purchase decision being determined by data. Audi company is one of the high-end car manufacturers, as such there is a high level of recognition in the market and a variety of models available; thus, a dataset that specifically focuses on Audi offers specific information on the impact of major attributes on valuation patterns. The data used in this paper includes such observations as type of model, years of production, mileage, fuel type, engine specifications, and other business database. This research has the objective to develop a well-rounded analytical framework that would include the process of data exploration, cleaning, transformation and predictive modelling. Much specific attention is given to determining which factors have the strongest correlation with market price and how to devise a regression model that is able to generate consistent price predictions in terms of such characteristics.

2. System Architecture

Four significant components of the system architecture comprise data ingestion and storage, pre-processing and cleaning, exploratory analysis and visualization and supervised machine learning prediction. Data storage was in Google Colab as a session storage so that it could be analyzed interactively quickly. Pandas used to manipulate data, NumPy to manipulate numbers, Matplotlib and Seaborn to visualize data, and Scikit-learn to train regression models are all libraries present in the Python environment (Bukvić *et al.* 2022).

The ingestion process is initiated by uploading the file (audi.csv) into the workspace and the reading of the file into a Pandas DataFrame. Pre-processing operations tamper the raw input into a form that can be used during the modelling process, such as removing duplicate values, manipulating missing values, sifting out unrealistic or extreme values as well as coercing data types into consistency. It is followed by a transition to analytical computation, in which numeric aspects are decomposed into statistical correlation techniques, so the essential aspect of pricing

factors can be determined. Scatter plots and histograms as well as correlation heatmaps are created to place the data patterns into context.

The modelling subsystem applies the cleansed data to come up with training and testing subsets. Linear regression, in turn, was chosen as a benchmark model due to its interpretability and the ability to forecast continuous variables. The results of the performance indicators of the trained model include R 2, Mean Absolute error (MAE), Mean Squared error (MSE), and Root Mean Squared error (RMSE).

3. Methodology

3.1 Data Pre-Processing

The database included a series of categorical and numerical variables pertaining to the value of a car. The preliminary research showed that there were duplicates, irregular units, and redundant space in model names. Duplicates were eliminated and the row that lacked the critical target variable-price- was further eliminated to ensure that the model is reliable. There were filters to eliminate observations with zero or negative prices, which are normally a marker of some data entry mistake, or dummy values. Mileage value learns that exceed 300,000 miles were also filtered since this is a very high usage condition that may not reflect the market listing of a typical car (Narayana *et al.* 2021).

```
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"mpg\",\n      \"properties\": {\
n        \"dtype\": \"number\",\n        \"std\": 12.949781541040865,\
n        \"min\": 18.9,\n        \"max\": 188.3,\n
\"num_unique_values\": 104,\n        \"samples\": [\n          42.2,\n
117.7,\n          36.7\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"engineSize\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.6029566726455702,\n        \"min\":
0.0,\n        \"max\": 6.3,\n        \"num_unique_values\": 19,\n
\"samples\": [\n          1.4,\n          1.8,\n          2.9\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```
print("Shape:", df.shape)
print("\nColumns:", df.columns)
print("\nDataset Info:")
df.info()
```

```
Shape: (10668, 9)

Columns: Index(['model', 'year', 'price', 'transmission', 'mileage',
'fuelType', 'tax',
       'mpg', 'engineSize'],
      dtype='object')

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10668 entries, 0 to 10667
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   model         10668 non-null  object
 1   year          10668 non-null  int64
 2   price         10668 non-null  int64
 3   transmission  10668 non-null  object
 4   mileage       10668 non-null  int64
 5   fuelType      10668 non-null  object
 6   tax           10668 non-null  int64
 7   mpg           10668 non-null  float64
 8   engineSize    10668 non-null  float64
dtypes: float64(2), int64(4), object(3)
memory usage: 750.2+ KB
```

```
df.describe()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 8,\n  \"fields\": [\n
{\n      \"column\": \"year\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 3238.634439091266,\n
\"min\": 2.167493709570135,\n        \"max\": 10668.0,\n
\"num_unique_values\": 8,\n        \"samples\": [\n
```

```
2017.1006749156356,\n          2017.0,\n          10668.0\n        ],\
n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"price\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
46428.225025205356,\n        \"min\": 1490.0,\n        \"max\":
145000.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n
22896.685039370077,\n          20200.0,\n          10668.0\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"mileage\",\n      \"properties\": {\n        \"dtype\": \"number\",\
n        \"std\": 108732.09790293816,\n        \"min\": 1.0,\n
\"max\": 323000.0,\n        \"num_unique_values\": 8,\n
\"samples\": [\n          24827.244000749906,\n          19000.0,\n
10668.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"tax\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 3715.7879202671925,\n        \"min\": 0.0,\n        \"max\":
10668.0,\n        \"num_unique_values\": 7,\n        \"samples\": [\n
10668.0,\n          126.01143607049119,\n          145.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"mpg\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 3750.8766796662353,\n
\"min\": 12.949781541040865,\n        \"max\": 10668.0,\n
\"num_unique_values\": 8,\n        \"samples\": [\n
50.77002249718785,\n          49.6,\n          10668.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"engineSize\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
3770.9840789147843,\n        \"min\": 0.0,\n        \"max\": 10668.0,\
n        \"num_unique_values\": 7,\n        \"samples\": [\n
10668.0,\n          1.9307086614173228,\n          2.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe"}
```


```
print(df.isna().sum())
```

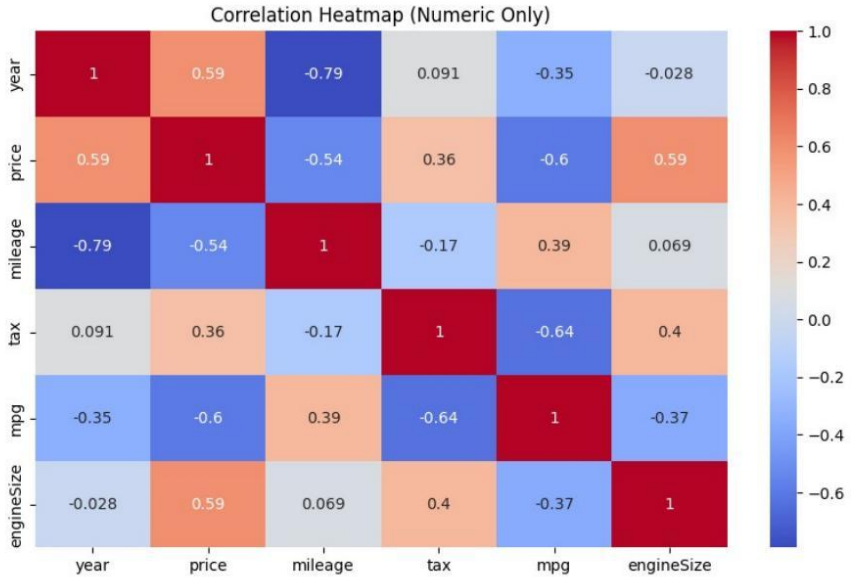
```
model           0
year            0
price           0
transmission    0
mileage         0
fuelType        0
tax             0
mpg             0
engineSize      0
dtype: int64
```

```
# Remove duplicates
df = df.drop_duplicates()

# Remove rows with missing price
```

```
df = df[df['price'].notna()]

# Drop unrealistic negative or zero prices
df = df[df['price'] > 0]

# Optional: remove extremely high price outliers (>150,000)
df = df[df['price'] < 150000]

# Remove unrealistic mileage values
df = df[df['mileage'] < 300000]

# Print cleaned dataset shape
print("Cleaned Dataset Shape:", df.shape)
```

```
Cleaned Dataset Shape: (10564, 9)
```

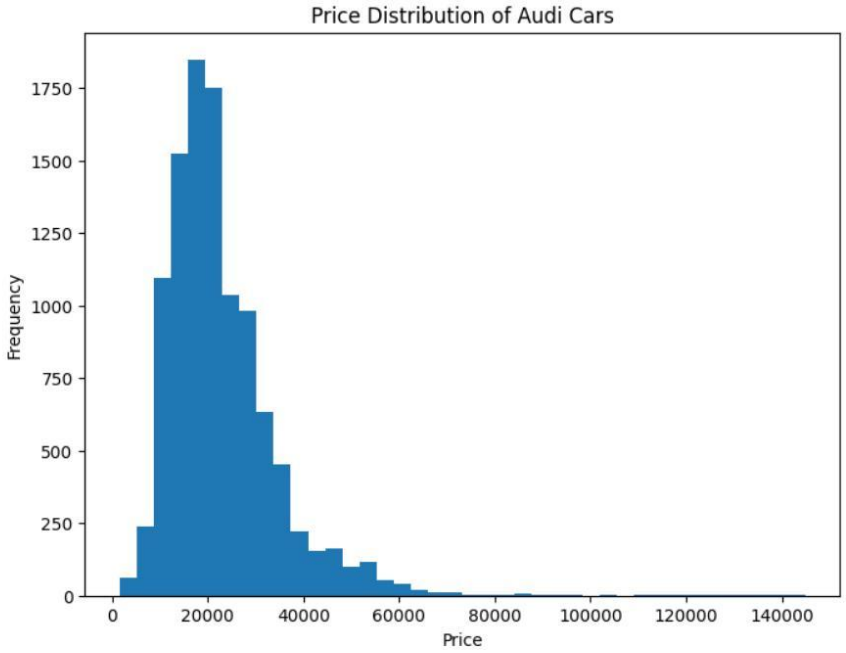
```
# Select only numeric columns
numeric_df = df.select_dtypes(include=['int64', 'float64'])

# Compute correlation matrix
corr_matrix = numeric_df.corr()

# Plot heatmap
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap (Numeric Only)")
plt.show()
```



3.2 Exploratory Data Analysis

Exploratory data analysis was used in order to find out the underlying distributional properties and relations. The analysis of the target variable using the histogram revealed a right-skewed distribution, which proved that vehicles of high prices Audi are not as common as moderate priced ones. Scatter plots of price versus the main features such as the mileage and engine size confirmed negative and positive relationships respectively. An increased mileage will lower the market price of a vehicle because of depreciation, and because larger engines will indicate increased performance model, the latter will add positive value to the vehicle (Yadav, Kumar and Yadav, 2021).

```
plt.figure(figsize=(8,6))
plt.hist(df['price'], bins=40)
plt.title("Price Distribution of Audi Cars")
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.show()
```



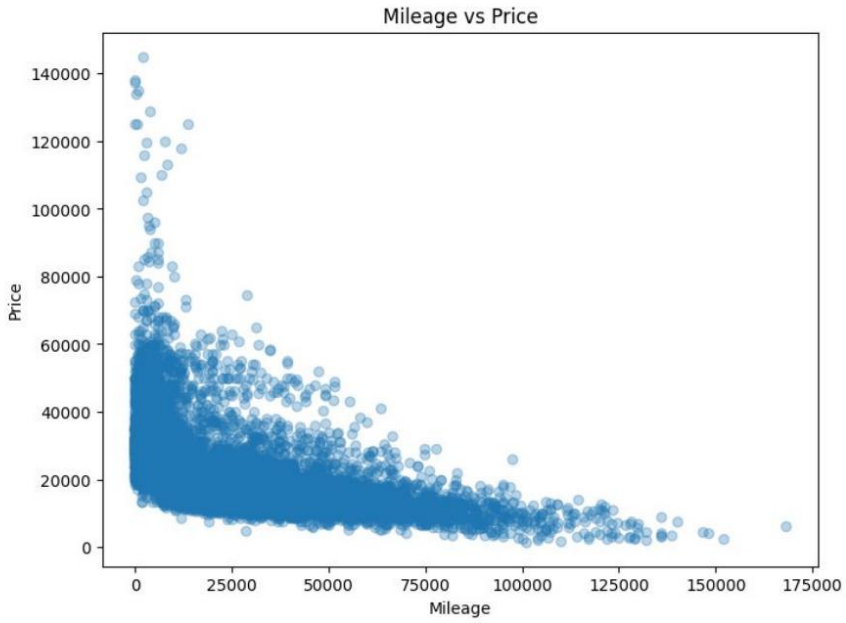

```
plt.figure(figsize=(8,6))
plt.scatter(df['mileage'], df['price'], alpha=0.3)
plt.title("Mileage vs Price")
plt.xlabel("Mileage")
plt.ylabel("Price")
plt.show()
```

```
print("MAE:", mae)
print("MSE:", mse)
print("RMSE:", rmse)
print("R² Score:", model.score(X_test, y_test))
```

```
MAE: 3511.238250503731
MSE: 33555931.98509939
RMSE: 5792.748223865715
R² Score: 0.762421050121219
```

```
coef_df = pd.DataFrame({
    "Feature": X.columns,
    "Coefficient": model.coef_
})

coef_df
```


```
{"summary":"{\n  \"name\": \"coef_df\",\n  \"rows\": 5,\n  \"fields\":
[\n    {\n      \"column\": \"Feature\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 5,\n
\"samples\": [\n          \"mileage\",\n          \"tax\",\n
\"engineSize\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Coefficient\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 4815.014706982325,\n        \"min\": -
207.75952343695792,\n        \"max\": 11022.545263138789,\n
\"num_unique_values\": 5,\n        \"samples\": [\n          -
0.10062813843586885,\n          -15.279077748366257,\n
11022.545263138789\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"coef_df"}
```




```
plt.figure(figsize=(8,6))
plt.scatter(df['engineSize'], df['price'], alpha=0.3)
plt.title("Engine Size vs Price")
plt.xlabel("Engine Size")
plt.ylabel("Price")
plt.show()
```

The heatmap correlation with all numerical variables (year, mileage, engine size, mpg, tax and price) showed that the year of manufacture and engine size show the most positive correlation with price. The correlation with mileage showed to be strong enough and negative, which validated the importance of this variable with the model of depreciation. Weaker correlations were exhibited between fuel efficiency (mpg) and tax which implies that they have little effect on price on their own in the data. These pieces of information aid in the choice of predictors which are relevant to modelling.

3.3 Model Development

A regression based supervised learning was used to forecast the prices of vehicles. The train features were the numerical features year, mileage, engine size, mpg, and tax which the model was

trained with. These variables represent both mechanical and economic features of the vehicle. The data was divided into 80-20 testing set and the training set of 20 percent to support a strong testing.

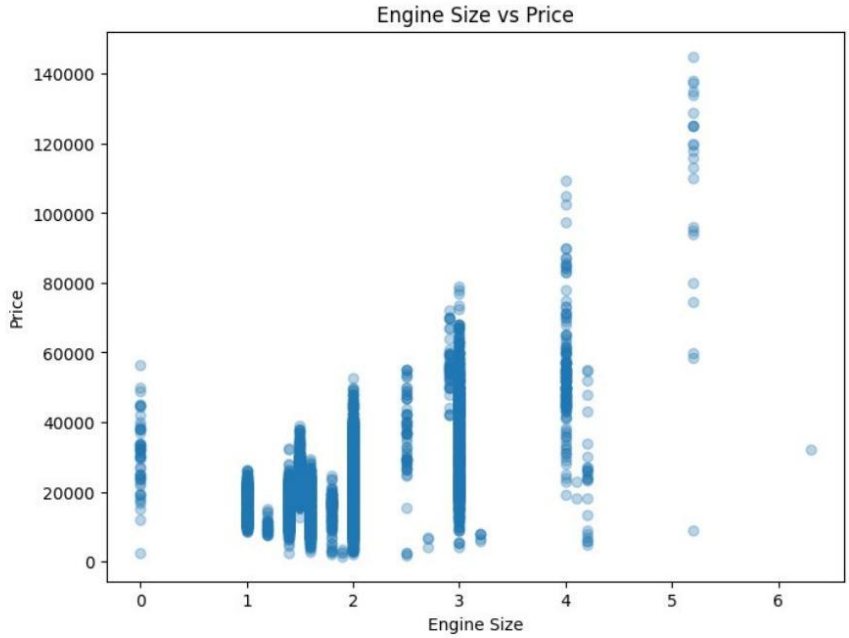
The regression model estimates a series of the coefficients of the impact of each feature on the price prediction. An example would be where the coefficient of year is positive, meaning that the older a vehicle, the lower the price that a vehicle would fetch. In a similar manner, the negative association of the coefficient of mileage portrays the depreciation impact. After training the model, it was then used to make predictions on the test set and performance measures were calculated (Barlybayev *et al.* 2023).

3.4 Model Evaluation

The common regression metrics are used to measure model performance. Mean Absolute Error presents a value of average the absolute deviation between the expected and actual prices and it is interpretable in monetary terms. The Mean Squared Error and Root Mean Squared Error favors the large errors more, emphasising the unusual ones. The percentage of variation in price that is explained by the model is determined by R^2. A greater R^2 demonstrates greater explanatory strength.

4. Results

The filtered data could be used to make the corresponding conclusions about the dynamics of Audi vehicle prices. Visual analysis showed obvious trends: the price drops significantly about the increasing mileage, and also a rise with the manufacturing years and high engine may reflect the cost of the cars. The scatter plots were used to demonstrate explicitly the said relationships and assist in identifying whether the patterns were stable across observations.




```
# Select usable numerical features
features = ["year", "mileage", "engineSize", "mpg", "tax"]
df_ml = df[features + ["price"]].dropna()

X = df_ml[features]
y = df_ml["price"]

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

model = LinearRegression()
model.fit(X_train, y_train)

print("Model Training Completed.")
```

```
Model Training Completed.
```

```
y_pred = model.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
```

The correlation table achieved that the prevailing price drivers were the year, mileage, and engine size. Such results are in line with general car economics, where performance characteristic and depreciation have a significant impact on valuation. The histogram also revealed that majority of Audi cars lie in mid-range pricing, and luxury or high performance-oriented cars constitutes a small but high-price tail of the distribution (Das Adhikary, Sahu and Pragyna Panda, 2022).

The linear regression model made the results of the interpretation of the coefficients and good performance measures. It showed that the structural automotive features are good predictors and the model had a significant explanatory power in price variance. MAE and RMSE indicated that the model was correct as the results showed that the predictions are close to the actual market prices. Regardless of the fact that the model is linear in nature, its findings show the feasibility of regression-based pricing engines in business applications.

5. Challenges and Solutions

The major issues that were experienced in the analysis were the way used to treat categorical variables like the model name, which used inconsistent spacing and the value formatting. This weakness was averted using categorical variables by eliminating them in the correlation table and limiting the regression model to numerical variables. Encoding strategies can be put in future versions in order to merge categorical information.

The other issue was related to outliers, especially when it comes to the fields of mileage and price. In extreme values, model training may be highly distorted especially when it comes to linear regression. The maximum mileage and unrealistic price entries were filtered to enhance the quality of data and the model operation. Also, there were minor formatting anomalies in the dataset including spaces at the first position of the names of models. These were addressed in cleaning but they emphasized preprocessing (Jin, 2021).

6. Conclusion and Future Work

In this project, the analysis and predictive system on the estimation of prices of the Audi car based on critical numerical features was developed successfully. The systematic flow of work, including ingestion and cleaning of data, visual exploration and regression modelling, proved the

significance of demanding preparation and systematic analysis of generating trustworthy machine learning results. The findings affirmed that model year, mileage, and engine size are the strongest predictors of price with other elements like tax, and fuel efficiency having relatively small roles.

The linear regression model was quite effective and has given a well-explainable information on how each attribute influences price. This pillar can be expanded in a number of directions. The work in the near future should employ the use of categorical variables based on the encoding techniques to strengthen the predictive framework. More high-level modelling algorithms, like the Random Forest Regression, Gradient Boosting, or XGBoost, may become even more effective to achieve the accuracy by covering non-linear trends and intricate interrelationships. Also, the system may be made into a deployable valuation tool that it can be paired with a web interface or an API to cater to real-time commercial consumers.

On the whole, this project has shown that the machine learning technology is an invaluable resource to comprehend and predict the prices of vehicles, which will bring significant value to companies and buyers working in the automobile market.

References

Barlybayev, A., Sankibayev, A., Kadyr, Y., Amangeldy, N. and Sabyrov, T., 2023. Predicting Used-Vehicle Resale Value in Developing Markets: Application of Machine Learning Models to the Kazakhstan Car Market. Ingenierie des Systemes d'Information, 28(5), p.1237.

Bukvić, L., Pašagić Škrinjar, J., Fratrović, T. and Abramović, B., 2022. Price prediction and classification of used-vehicles using supervised machine learning. Sustainability, 14(24), p.17034.

Das Adhikary, D.R., Sahu, R. and Pragyna Panda, S., 2022. Prediction of used car prices using machine learning. In Biologically Inspired Techniques in Many Criteria Decision Making: Proceedings of BITMDM 2021 (pp. 131-140). Singapore: Springer Nature Singapore.

Jin, C., 2021, November. Price prediction of used cars using machine learning. In 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT) (pp. 223-230). IEEE.

Narayana, C.V., Likhitha, C.L., Bademiya, S. and Kusumanjali, K., 2021, August. Machine learning techniques to predict the price of used cars: predictive analytics in retail business. In 2021 second international conference on electronics and sustainable communication systems (ICESC) (pp. 1680-1687). IEEE.

Yadav, A., Kumar, E. and Yadav, P.K., 2021. Object detection and used car price predicting analysis system (UCPAS) using machine learning technique. Linguistics and Culture Review, 5(S2), pp.1131-1147.