1. Develop RF prediction model for consumption.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, cross_val_score, learning_curve,
GridSearchCV
from sklearn.inspection import permutation_importance
import statistics

# Read the data
df = pd.read_excel("/Users/felipemarques/Documents/GitHub/ucw/predictive-
analytics/Midterm Exam/Midterm-Dataset (1).xlsx")


#Split the date into quarter and year and make it a Numeric Date based in quartiles
for the year
df[['Quarter', 'Year']] = df['Date'].str.split(', ', expand=True)
quarter_mapping = {'Q1': 1, 'Q2': 2, 'Q3': 3, 'Q4': 4}
df['Quarter'] = df['Quarter'].map(quarter_mapping)
df['Year'] = df['Year'].astype(int)
df['NumericDate'] = df['Year'] + (df['Quarter'] - 1) / 4

# Split the data into X and y
y = df[ "Consumption"]
x = df[[ "NumericDate", "Income", "Job type", "Location"]]


# One-hot encode the data to make it usable by the model (strings to numbers)
x_encoded = pd.get_dummies(x, columns=["Location"])

# Split the data into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(x_encoded, y, test_size=0.2,
random_state=42)

forest_model = RandomForestRegressor(random_state=42, max_depth=7, max_features=7,
n_estimators=100)
forest_model.fit(x_train, y_train)

# Predict the test data
y_pred = forest_model.predict(x_test)
```

2. Rationalize the chosen structure of the model.

```
# Without numeric date (or any standard date format), the model would not be able to
predict the consumption well as it did. It is known that consumption of good and
services are subjected to seasonal variance, therefore, it is paramount to have the
```

```
date. This relevance for date can be interpreted by the Feature and Permutation
Importance, which was highly ranked for 'NumericDate'. The model performance almost
perfectly, with the highest number of R-squared (100%) and was also able to predict
the consumption for the test data with a very low Mean Absolute Error (83.86 for
absolute and 0.27 for percentage). The best hyperparameter for the
RandomForestRegressor was 7 features and 100 estimators in a maximum depth of 7.
```

3. What is the predicted consumption if the disposal income is $33,000 (use mean/mode for other variables if used)?

```
The predicted consumption is $31,773.30
```

4. Simulate the model variables

```
Simulate new input values for the predictor variables
simulate_data = pd.DataFrame({
    'NumericDate': [2010.25, 2011.50, 2012.75, 2015.25, 2015.75, 2016.75],
    'Income': [5000, 16000, 27000, 32000, 45000, 80000],
    'Job type': [1, 1, 0, 0, 1, 1],
    'Location_Canada ': [1, 0, 0, 0, 1, 0],
    'Location_Europe ': [0, 1, 0, 1, 0, 0],
    'Location_USA': [0, 0, 1, 0, 0, 1]
```

5. Comment on the predicted values

```
# Although we had a great performance for Random Forest (near perfect R-squared), the
model does not seem to be very accurate for the simulated data. The predicted
consumption for the simulated data with low income (lower than any other appearance in
the dataset) is bigger than the income, breaking fundaments of comsumption vs. incone
(you won't be able to spend more than you earn, without a loan). This happened in
almost all attempts of simulations (in exclusion to income is 45k or 80k). Althogh
some might think that this is the case of overfitting, this is not true because both
train and test data showed excellent performance (high r-squared and low mape). This
is an inherited limitation of Random Forest (also Decision Tree) models, that won't
perform very well if new sample when compared to the train data has considerable
magninute differences.
```

6. What are the limitations of the model?

```
# The model is limited to the data that was used to train it. If the data is not
representative of the population, the model will not be able to predict well. Also,
the model is limited to the variables that were used to train it. If there are other
variables that are important to predict the consumption, the model will not be able to
predict well. If new observations has a considerable difference in magnitude from the
training data, the model will not be able to predict well.
```