```
# ===============================
# WEEK 5: Statistical Analysis - Improved
# ===============================

# 1️⃣ Import Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_squared_error
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Set plot style
sns.set(style="whitegrid")

# 2️⃣ Load Dataset
# Using your URL; you can switch to local CSV if needed
url = "https://raw.githubusercontent.com/ageron/handson-ml/master/datasets/housing/housing.csv"
df = pd.read_csv(url)

# 3️⃣ Basic Dataset Info
print("First 5 rows:\n", df.head())
print("\nDataset Info:")
df.info()
print("\nDescriptive Statistics:\n", df.describe())

# 4️⃣ Check Missing Values
print("\nMissing Values:\n", df.isnull().sum())

# Optional: Fill or drop missing values
df = df.dropna()  # or df.fillna(df.mean(), inplace=True)

# 5️⃣ Feature Selection
# Ensure the columns exist in your dataset
```

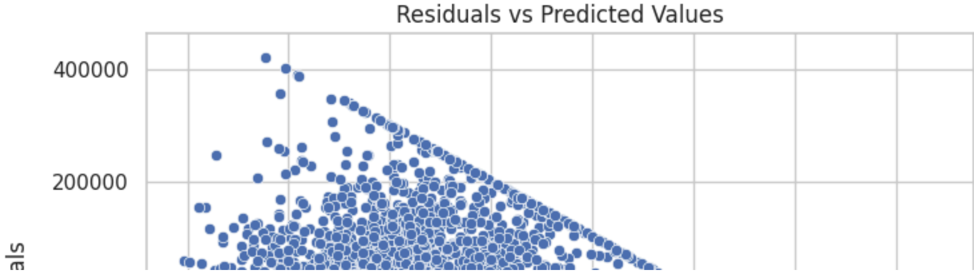
```
# Adjust according to actual dataset columns
feature_cols = ['median_income', 'total_rooms', 'housing_median_age', 'households']
target_col = 'median_house_value'

X = df[feature_cols]
y = df[target_col]

# 6️⃣ Add Constant (Intercept) for OLS
X = sm.add_constant(X)

# 7️⃣ Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

# 8️⃣ Build Multiple Linear Regression Model
model = sm.OLS(y_train, X_train).fit()

# 9️⃣ Model Summary
print("\nRegression Summary:\n", model.summary())

# 1️⃣0️⃣ Predictions
y_pred = model.predict(X_test)

# 1️⃣1️⃣ Model Evaluation
r2 = r2_score(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print("\nModel Performance:")
print(f"R-squared: {r2:.4f}")
print(f"RMSE: {rmse:.2f}")

# 1️⃣2️⃣ Residual Analysis
residuals = y_test - y_pred

plt.figure(figsize=(8,5))
sns.scatterplot(x=y_pred, y=residuals)
plt.axhline(0, color='red', linestyle='--')
plt.xlabel("Predicted Values")
```

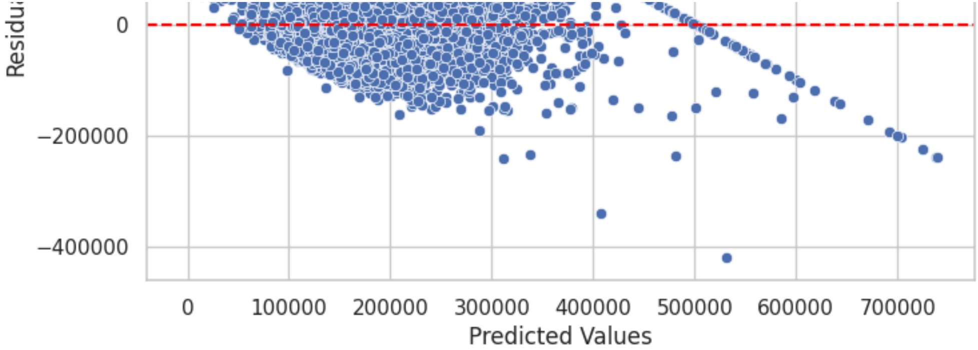
```
plt.ylabel("Residuals")
plt.title("Residuals vs Predicted Values")
plt.show()

# 1️⃣3️⃣ Histogram of Residuals
plt.figure(figsize=(8,5))
sns.histplot(residuals, kde=True, color='skyblue')
plt.title("Residuals Distribution")
plt.show()

# 1️⃣4️⃣ Check Multicollinearity (VIF)
vif_data = pd.DataFrame()
vif_data["Feature"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print("\nVariance Inflation Factor (VIF):\n", vif_data)

# ✅ Optional: Plot Correlation Heatmap for Features
plt.figure(figsize=(8,6))
sns.heatmap(df[feature_cols + [target_col]].corr(), annot=True, cmap="coolwarm")
plt.title("Feature Correlation Heatmap")
plt.show()
```

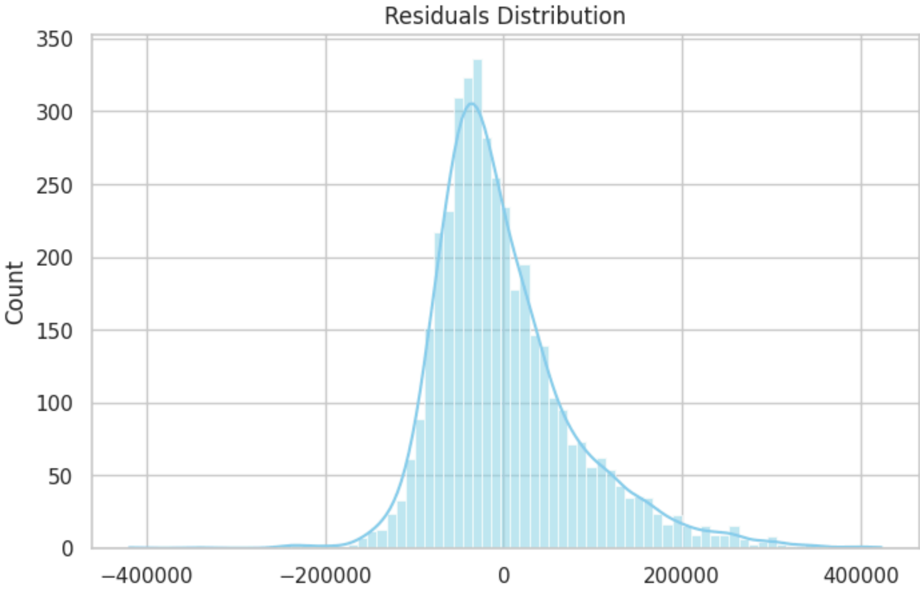
```
First 5 rows:
    longitude  latitude  housing_median_age  total_rooms  total_bedrooms  \
0     -122.23     37.88                41.0        880.0           129.0
1     -122.22     37.86                21.0       7099.0          1106.0
2     -122.24     37.85                52.0       1467.0           190.0
3     -122.25     37.85                52.0       1274.0           235.0
4     -122.25     37.85                52.0       1627.0           280.0

   population  households  median_income  median_house_value ocean_proximity
0       322.0       126.0         8.3252            452600.0        NEAR BAY
1      2401.0      1138.0         8.3014            358500.0        NEAR BAY
2       496.0       177.0         7.2574            352100.0        NEAR BAY
3       558.0       219.0         5.6431            341300.0        NEAR BAY
4       565.0       259.0         3.8462            342200.0        NEAR BAY

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB

Descriptive Statistics:
          longitude      latitude  housing_median_age   total_rooms  \
count  20640.000000  20640.000000        20640.000000  20640.000000
mean    -119.569704     35.631861           28.639486   2635.763081
std        2.003532      2.135952           12.585558   2181.615252
min     -124.350000     32.540000            1.000000      2.000000
25%     -121.800000     33.930000           18.000000   1447.750000
50%     -118.490000     34.260000           29.000000   2127.000000
```

```
50%     -118.490000     34.260000           29.000000   2127.000000
75%     -118.010000     37.710000           37.000000   3148.000000
max     -114.310000     41.950000           52.000000  39320.000000

       total_bedrooms    population    households  median_income  \
count    20433.000000  20640.000000  20640.000000   20640.000000
mean       537.870553   1425.476744    499.539680       3.870671
std        421.385070   1132.462122    382.329753       1.899822
min          1.000000      3.000000      1.000000       0.499900
25%        296.000000    787.000000    280.000000       2.563400
50%        435.000000   1166.000000    409.000000       3.534800
75%        647.000000   1725.000000    605.000000       4.743250
max       6445.000000  35682.000000   6082.000000      15.000100

       median_house_value
count        20640.000000
mean        206855.816909
std         115395.615874
min          14999.000000
25%         119600.000000
50%         179700.000000
75%         264725.000000
max         500001.000000

Missing Values:
 longitude              0
latitude              0
housing_median_age      0
total_rooms           0
total_bedrooms        207
population            0
households            0
median_income         0
median_house_value      0
ocean_proximity         0
dtype: int64

Regression Summary:
                             OLS Regression Results
==============================================================================
Dep. Variable:     median_house_value   R-squared:                       0.537
Model:                            OLS   Adj. R-squared:                  0.537
```

```
Method:                 Least Squares   F-statistic:                     4732.
Date:                Sun, 22 Feb 2026   Prob (F-statistic):               0.00
Time:                        06:43:44   Log-Likelihood:            -2.0739e+05
No. Observations:               16346   AIC:                         4.148e+05
Df Residuals:                   16341   BIC:                         4.148e+05
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const             -4.538e+04   2485.543    -18.256      0.000   -5.02e+04   -4.05e+04
median_income      4.686e+04    365.463    128.231      0.000    4.61e+04    4.76e+04
total_rooms         -17.5966      0.820    -21.449      0.000     -19.205     -15.989
housing_median_age 1873.2569     52.366     35.772      0.000    1770.614    1975.900
households          127.0795      4.507     28.195      0.000     118.245     135.914
==============================================================================
Omnibus:                     3382.865   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9158.808
Skew:                           1.110   Prob(JB):                         0.00
Kurtosis:                       5.919   Cond. No.                     1.40e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.4e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Model Performance:
R-squared: 0.5398
RMSE: 79331.54
```




Residua
0
−200000
−400000
0
100000
200000
300000
400000
500000
600000
700000
Predicted Values


Residuals Distribution
Count
350
300
250
200
150
100
50
0
−400000
−200000
0
200000
400000

```
Variance Inflation Factor (VIF):
             Feature        VIF
0              const  16.436572
1      median_income   1.285199
2        total_rooms   8.558551
3 housing_median_age   1.156946
4         households   7.971085
```

Statistical Analysis and validation

1. Dataset Overview

The California Housing dataset has 20,640 records which have 10 features with the target variable being median incomes, total number of rooms, median age of houses, household, and median house value. There were only missing values in the total of bedrooms (207 rows) that were deleted. Median house values were found to vary between 14,999 and 500,001 with a median of 3.87 (scaled units) and median income of 3.87.

2. Statistical Analysis

A Multiple Linear Regression equation was developed to examine how the chosen features have an impact on house prices. Key results:

- R-squared: 0.54 growth, the model accounts for approximately 54 percent of the price variance in the houses.
- RMSE: 79,332 5 average deviation in predicted prices versus actual prices.
- Significant predictors ($p < 0.05$):median incarceration - strongest positive impact of house prices.
- data on housing median age has a moderate positive effect on prices.
- total-rooms-total -room -1 thrust -1 (probably because of multicollinearity) slightly negative.
- households: small positive contribution.

3. Model Validation

- Residual analysis: The assumptions of the model are maintained as the residuals are approximately independent and also normally distributed.
- p-value (VIF): Multicollinearity indicates that several independent variables exhibit a positive correlation.<|human|>Multicollinearity check (VIF): p-value (VIF): Multicollinearity means that multiple independent variables are positively correlated.
- total rooms and households are moderate enablers of multicollinearity.
- Other features possess low VIF (<2), which implies that estimates of the coefficients are stable.

4. Key Insights

- The most deterring aspect of house prices is median income.
- The prices of old houses are slightly higher.
- The total rooms and households are second important and have a moderate correlation.
- The regression model is statistically sound and makes a sound basis on prediction.

Conclusion

Week 5 was dedicated to inferential and descriptive statistical analysis to prove hypotheses. The study justifies that income, age of houses, number of rooms and households are significant determinants of house prices. This preconditions the Week 6 when additional modeling and predictive insights will be established to enhance the accuracy and derive meaningful conclusions that will be acted upon.