

# CoffeeSales

September 11, 2025

These analyses explore the variables that are associated with revenue sold from coffee.

Analysis used: stepwise regression

Created on: August 22, 2025 Created by: Claudia Claudia

## Import Libraries

```
[2]: import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
from statsmodels.stats.outliers_influence import variance_inflation_factor
import matplotlib.pyplot as plt
import itertools
```

## Read in the data

```
[3]: address = '/Users/claudiaclinchard/Desktop/Projects/Coffee_sales/Coffee_sales.
      ↪xlsx'

coffee = pd.read_excel(address, header=int(1))
coffee.columns=['date', 'datetime', 'hour_of_day', 'cash_type', 'card',
      ↪'money', 'coffee_name',
      'time_of_day', 'weekday', 'month_name', 'weekdaysort',
      ↪'monthsort']

coffee[:5]
```

```
[3]:      date      datetime  hour_of_day  cash_type  \
0 2024-03-01 2024-03-01 12:19:22.539         12    card
1 2024-03-01 2024-03-01 12:20:18.089         12    card
2 2024-03-01 2024-03-01 13:46:33.006         13    card
3 2024-03-01 2024-03-01 13:48:14.626         13    card
4 2024-03-01 2024-03-01 15:39:47.726         15    card

      card  money      coffee_name  time_of_day  weekday  \
0  ANON-0000-0000-0002    38.7    Hot Chocolate    Afternoon    Fri
1  ANON-0000-0000-0002    38.7    Hot Chocolate    Afternoon    Fri
2  ANON-0000-0000-0003    28.9    Americano    Afternoon    Fri
3  ANON-0000-0000-0004    38.7         Latte    Afternoon    Fri
```

4	ANON-0000-0000-0005	33.8	Americano with Milk	Afternoon	Fri
---	---------------------	------	---------------------	-----------	-----

	month_name	weekdaysort	monthsort
0	Mar	5	3
1	Mar	5	3
2	Mar	5	3
3	Mar	5	3
4	Mar	5	3

```
[4]: X = coffee[['hour_of_day', 'cash_type', 'coffee_name', 'weekday', 'month_name']]
      y = coffee[['money']]
```

```
[76]: for col in coffee.select_dtypes(include="object").columns:
      coffee[col] = coffee[col].astype("category")

# Descriptives
print(coffee.head())
print(coffee.describe(include="all"))
print(coffee.dtypes)
print({col: coffee[col].nunique() for col in coffee.
      ↪select_dtypes(include="category").columns})

coffee["hour_of_day"] = coffee["hour_of_day"].astype("category")
```

	date	datetime	hour_of_day	cash_type	\
0	2024-03-01	2024-03-01 12:19:22.539	12	card	
1	2024-03-01	2024-03-01 12:20:18.089	12	card	
2	2024-03-01	2024-03-01 13:46:33.006	13	card	
3	2024-03-01	2024-03-01 13:48:14.626	13	card	
4	2024-03-01	2024-03-01 15:39:47.726	15	card	

	card	money	coffee_name	time_of_day	weekday	\
0	ANON-0000-0000-0002	38.7	Hot Chocolate	Afternoon	Fri	
1	ANON-0000-0000-0002	38.7	Hot Chocolate	Afternoon	Fri	
2	ANON-0000-0000-0003	28.9	Americano	Afternoon	Fri	
3	ANON-0000-0000-0004	38.7	Latte	Afternoon	Fri	
4	ANON-0000-0000-0005	33.8	Americano with Milk	Afternoon	Fri	

	month_name	weekdaysort	monthsort
0	Mar	5	3
1	Mar	5	3
2	Mar	5	3
3	Mar	5	3
4	Mar	5	3

	date	datetime	\
count	3635	3635	
unique	NaN	NaN	
top	NaN	NaN	

freq		NaN		NaN
mean	2024-09-30 13:20:36.973865472	2024-10-01 04:00:09.484163584		
min	2024-03-01 00:00:00	2024-03-01 12:19:22.539000		
25%	2024-07-03 00:00:00	2024-07-03 16:54:57.165000192		
50%	2024-10-07 00:00:00	2024-10-07 08:33:36.423000064		
75%	2025-01-08 00:00:00	2025-01-08 08:20:38.580500224		
max	2025-03-23 00:00:00	2025-03-23 18:11:38.635000		
std		NaN		NaN

	hour_of_day	cash_type	card	money	\
count	3635.000000	3635	3546	3635.000000	
unique	NaN	2	1316	NaN	
top	NaN	card	ANON-0000-0000-0012	NaN	
freq	NaN	3546	129	NaN	
mean	14.168088	NaN	NaN	31.744946	
min	6.000000	NaN	NaN	18.120000	
25%	10.000000	NaN	NaN	27.920000	
50%	14.000000	NaN	NaN	32.820000	
75%	18.000000	NaN	NaN	35.760000	
max	22.000000	NaN	NaN	40.000000	
std	4.227771	NaN	NaN	4.919250	

	coffee_name	time_of_day	weekday	month_name	weekdaysort	\
count	3635	3635	3635	3635	3635.000000	
unique	8	3	7	12	NaN	
top	Americano with Milk	Afternoon	Tue	Mar	NaN	
freq	824	1231	585	524	NaN	
mean	NaN	NaN	NaN	NaN	3.847593	
min	NaN	NaN	NaN	NaN	1.000000	
25%	NaN	NaN	NaN	NaN	2.000000	
50%	NaN	NaN	NaN	NaN	4.000000	
75%	NaN	NaN	NaN	NaN	6.000000	
max	NaN	NaN	NaN	NaN	7.000000	
std	NaN	NaN	NaN	NaN	1.976163	

	monthsort
count	3635.000000
unique	NaN
top	NaN
freq	NaN
mean	6.395598
min	1.000000
25%	3.000000
50%	6.000000
75%	10.000000
max	12.000000
std	3.480709
date	datetime64[ns]

```

datetime        datetime64[ns]
hour_of_day      int64
cash_type        category
card             category
money           float64
coffee_name      category
time_of_day      category
weekday          category
month_name       category
weekdaysort     int64
monthsort        int64
dtype: object
{'cash_type': 2, 'card': 1316, 'coffee_name': 8, 'time_of_day': 3, 'weekday': 7,
'month_name': 12}

```

[77]: *# Barplots of proportions*

```

#Coffee Type
counts = coffee['coffee_name'].value_counts(normalize=True)
plt.bar(counts.index, counts.values, color = "#7B3F00")
plt.ylabel("Proportion")
plt.xlabel("Coffee Type")
plt.xticks(rotation=45)
plt.show()

#Hour of Day
counts = coffee['hour_of_day'].value_counts(normalize=True)
plt.bar(counts.index, counts.values, color = "#A52A2A")
plt.ylabel("Proportion")
plt.xlabel("Hour of Day")
plt.xticks(rotation=45)
plt.show()

#Cash Type
counts = coffee['cash_type'].value_counts(normalize=True)
plt.bar(counts.index, counts.values, color = "#C19A6B")
plt.ylabel("Proportion")
plt.xlabel("Cash Type")
plt.xticks(rotation=45)
plt.show()

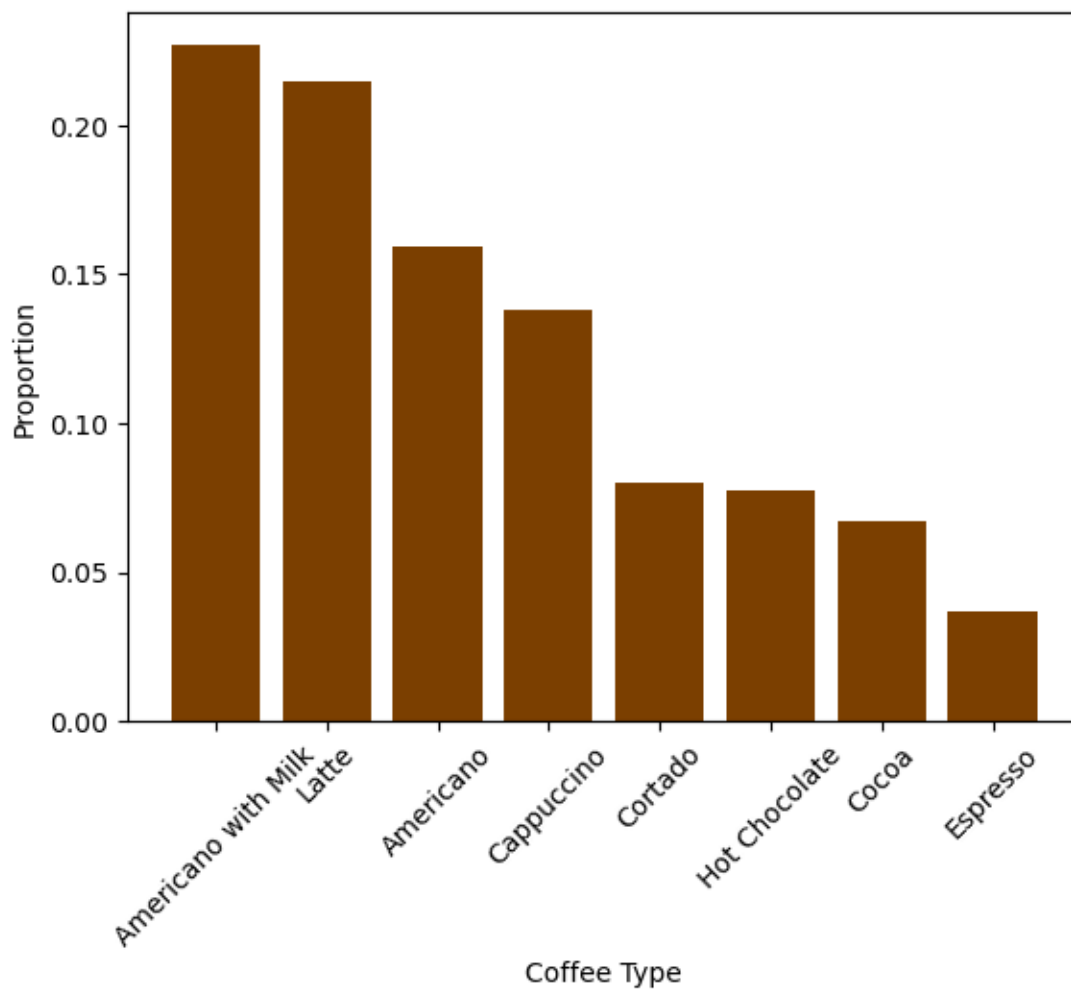
#Time of Day
counts = coffee['time_of_day'].value_counts(normalize=True)
plt.bar(counts.index, counts.values, color = "#7B3F00")
plt.ylabel("Proportion")
plt.xlabel("Time of Day")
plt.xticks(rotation=45)

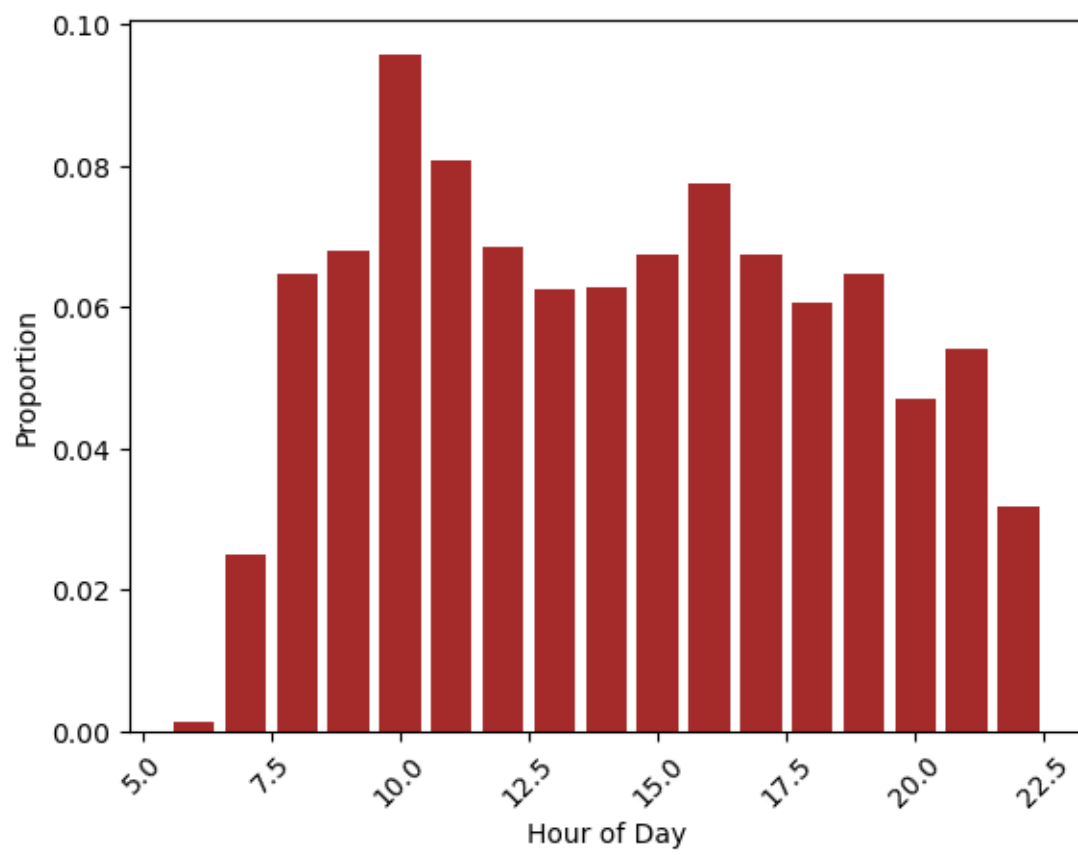
```

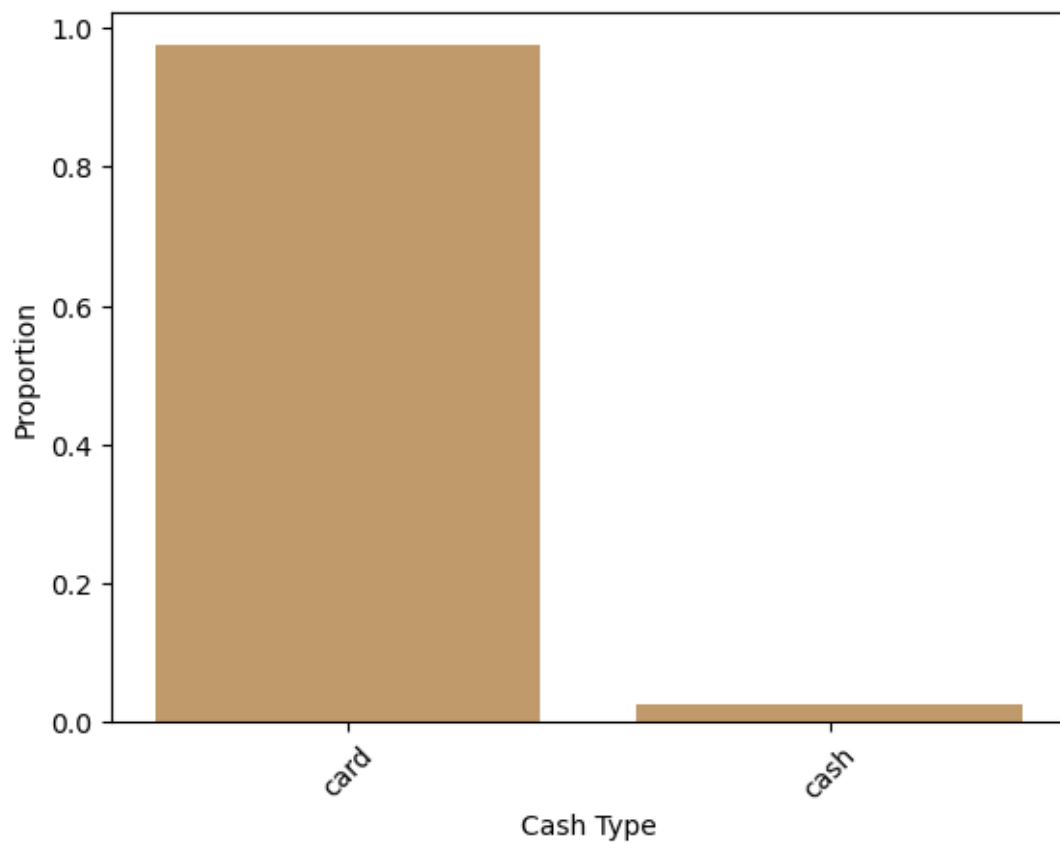
```
plt.show()

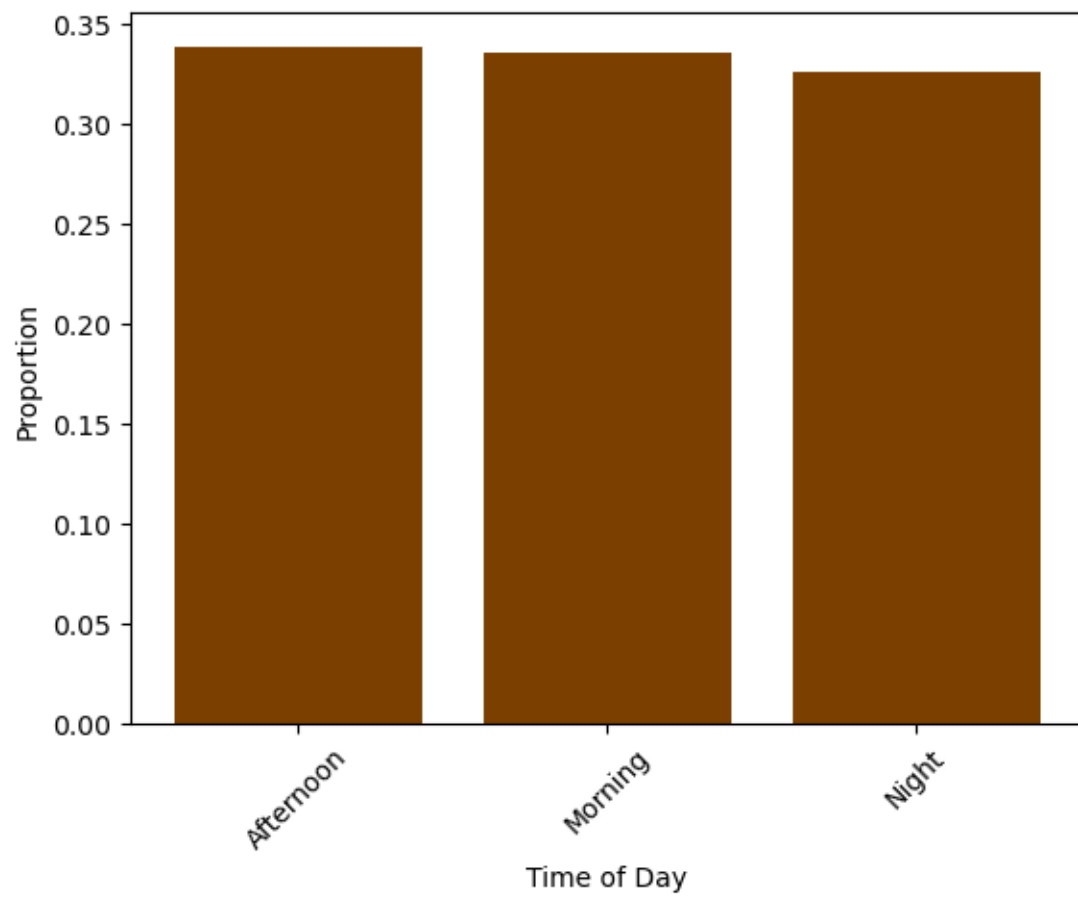
#Weekday
counts = coffee['weekday'].value_counts(normalize=True)
plt.bar(counts.index, counts.values, color = "#A52A2A")
plt.ylabel("Proportion")
plt.xlabel("Weekday")
plt.xticks(rotation=45)
plt.show()

#Month
counts = coffee['month_name'].value_counts(normalize=True)
plt.bar(counts.index, counts.values, color = "#A52A2A")
plt.ylabel("Proportion")
plt.xlabel("Month")
plt.xticks(rotation=45)
plt.show()
```

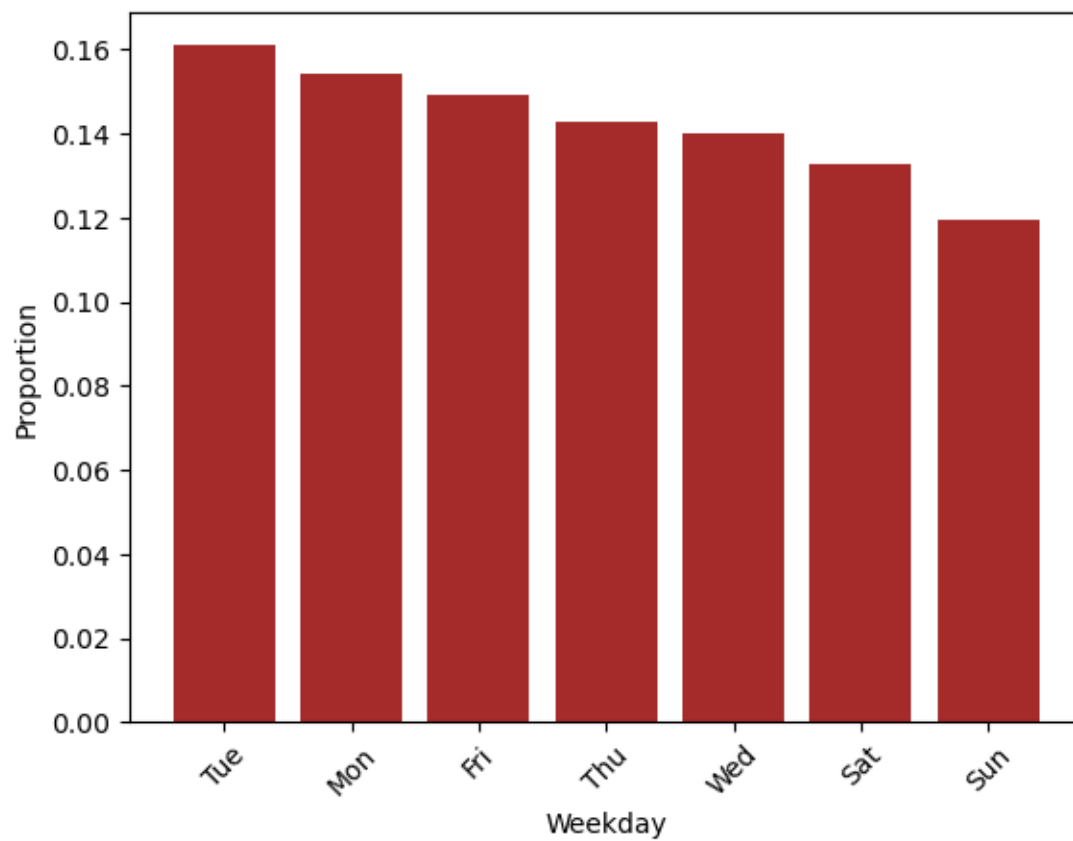


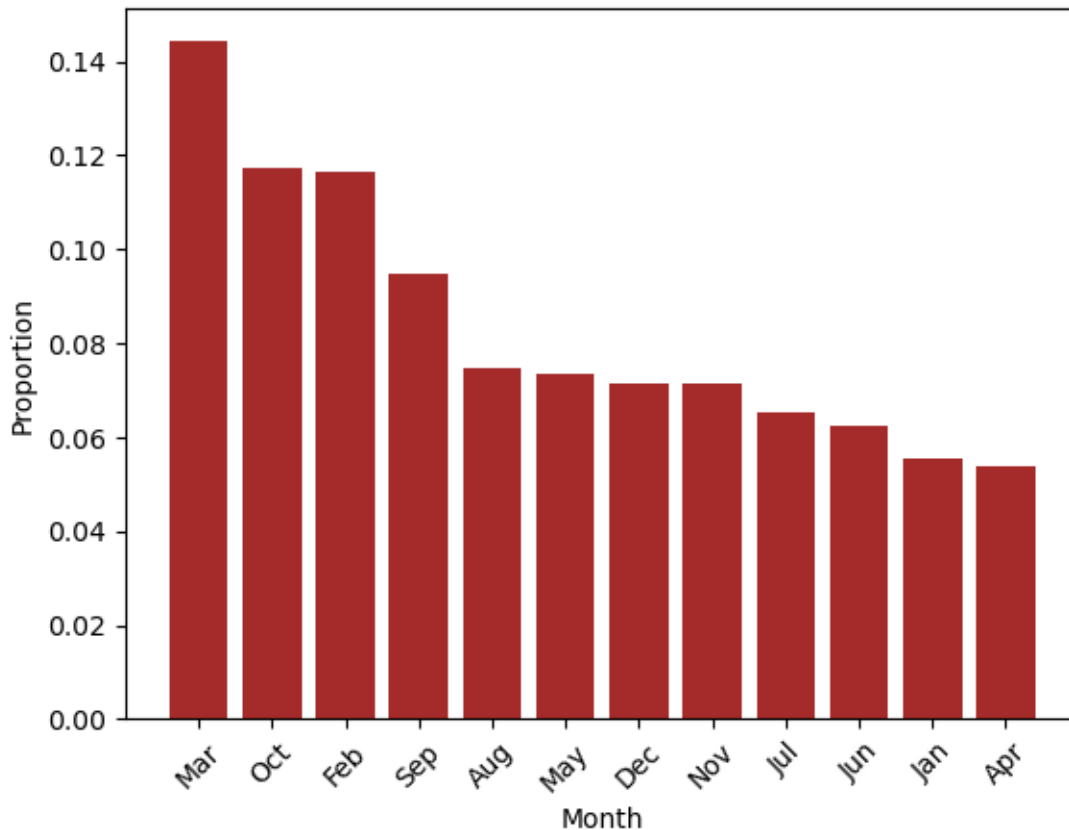












```
[12]: #Full Model
full_formula = "money ~ hour_of_day + cash_type + coffee_name + time_of_day +_
    ↪weekday + month_name"
full_mod = smf.ols(full_formula, data=coffee).fit()
print(full_mod.summary())
print("AIC full model:", full_mod.aic)
```

#### OLS Regression Results

```
=====
Dep. Variable:          money    R-squared:                0.978
Model:                  OLS      Adj. R-squared:           0.978
Method:                 Least Squares    F-statistic:             5668.
Date:                   Thu, 11 Sep 2025    Prob (F-statistic):       0.00
Time:                   12:14:50    Log-Likelihood:          -4029.5
No. Observations:      3635    AIC:                     8117.
Df Residuals:          3606    BIC:                     8297.
Df Model:               28
Covariance Type:       nonrobust
=====
```

		coef	std err	t	P> t
[0.025	0.975]				
-----					
Intercept		28.3809	0.135	210.487	0.000
28.117	28.645				
cash_type[T.cash]		1.9813	0.082	24.071	0.000
1.820	2.143				
coffee_name[T.Americano with Milk]		5.0052	0.041	122.739	0.000
4.925	5.085				
coffee_name[T.Cappuccino]		9.8824	0.046	215.750	0.000
9.793	9.972				
coffee_name[T.Cocoa]		9.7770	0.057	170.596	0.000
9.665	9.889				
coffee_name[T.Cortado]		0.2853	0.054	5.287	0.000
0.179	0.391				
coffee_name[T.Espresso]		-4.7084	0.071	-66.313	0.000
-4.848	-4.569				
coffee_name[T.Hot Chocolate]		9.9176	0.055	179.955	0.000
9.810	10.026				
coffee_name[T.Latte]		9.8856	0.041	238.857	0.000
9.804	9.967				
time_of_day[T.Morning]		-0.1236	0.050	-2.476	0.013
-0.221	-0.026				
time_of_day[T.Night]		-0.0503	0.053	-0.957	0.339
-0.153	0.053				
weekday[T.Mon]		-0.0098	0.045	-0.220	0.826
-0.097	0.078				
weekday[T.Sat]		0.0442	0.047	0.950	0.342
-0.047	0.135				
weekday[T.Sun]		-0.0030	0.048	-0.062	0.951
-0.097	0.091				
weekday[T.Thu]		-0.0476	0.045	-1.047	0.295
-0.137	0.041				
weekday[T.Tue]		-0.0347	0.044	-0.787	0.432
-0.121	0.052				
weekday[T.Wed]		-0.0343	0.046	-0.752	0.452
-0.124	0.055				
month_name[T.Aug]		-5.4470	0.071	-77.253	0.000
-5.585	-5.309				
month_name[T.Dec]		-2.5134	0.071	-35.376	0.000
-2.653	-2.374				
month_name[T.Feb]		-2.4756	0.065	-37.941	0.000
-2.604	-2.348				
month_name[T.Jan]		-2.5138	0.075	-33.456	0.000
-2.661	-2.367				
month_name[T.Jul]		-4.7882	0.073	-65.859	0.000
-4.931	-4.646				

month_name[T. Jun]	-0.5743	0.073	-7.859	0.000
-0.718	-0.431			
month_name[T. Mar]	-1.3813	0.062	-22.133	0.000
-1.504	-1.259			
month_name[T. May]	-0.6166	0.070	-8.861	0.000
-0.753	-0.480			
month_name[T. Nov]	-2.5000	0.071	-34.995	0.000
-2.640	-2.360			
month_name[T. Oct]	-2.4977	0.065	-38.289	0.000
-2.626	-2.370			
month_name[T. Sep]	-5.1645	0.067	-76.514	0.000
-5.297	-5.032			
hour_of_day	0.0042	0.008	0.494	0.621
-0.012	0.021			

---

Omnibus:	1996.907	Durbin-Watson:	0.090
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19650.873
Skew:	2.447	Prob(JB):	0.00
Kurtosis:	13.286	Cond. No.	237.

---

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

AIC full model: 8117.099077791354

```
[14]: #Examine VIF for multicollinearity
X=full_mod.model.exog
features=full_mod.model.exog_names
vif_data = pd.DataFrame()
vif_data["feature"] = features
vif_data["VIF"] = [variance_inflation_factor(X, i) for i in range(X.shape[1])]
vif_data = vif_data[vif_data["feature"] != "Intercept"]
print(vif_data)
```

	feature	VIF
1	cash_type[T.cash]	1.085652
2	coffee_name[T.Americano with Milk]	1.955621
3	coffee_name[T.Cappuccino]	1.672578
4	coffee_name[T.Cocoa]	1.374555
5	coffee_name[T.Cortado]	1.442946
6	coffee_name[T.Espresso]	1.200778
7	coffee_name[T.Hot Chocolate]	1.458122
8	coffee_name[T.Latte]	1.938477
9	time_of_day[T.Morning]	3.725417
10	time_of_day[T.Night]	4.066860
11	weekday[T.Mon]	1.736439
12	weekday[T.Sat]	1.669770

```

13             weekday[T.Sun]  1.620571
14             weekday[T.Thu]  1.696247
15             weekday[T.Tue]  1.758006
16             weekday[T.Wed]  1.682200
17             month_name[T.Aug] 2.308853
18             month_name[T.Dec] 2.240928
19             month_name[T.Feb] 2.936855
20             month_name[T.Jan] 1.978495
21             month_name[T.Jul] 2.161284
22             month_name[T.Jun] 2.097183
23             month_name[T.Mar] 3.223839
24             month_name[T.May] 2.210337
25             month_name[T.Nov] 2.265712
26             month_name[T.Oct] 2.953383
27             month_name[T.Sep] 2.618734
28             hour_of_day    8.615598

```

```

[15]: #Remove hour_of_day
full_formula2 = "money ~ cash_type + coffee_name + time_of_day + weekday +
↪month_name"
full_mod2 = smf.ols(full_formula2, data=coffee).fit()
print(full_mod2.summary())
print("AIC full model:", full_mod2.aic)

```

#### OLS Regression Results

```

=====
Dep. Variable:          money    R-squared:                0.978
Model:                  OLS      Adj. R-squared:            0.978
Method:                 Least Squares    F-statistic:          5879.
Date:                   Thu, 11 Sep 2025    Prob (F-statistic):    0.00
Time:                   12:15:08    Log-Likelihood:        -4029.7
No. Observations:       3635    AIC:                   8115.
Df Residuals:           3607    BIC:                   8289.
Df Model:                27
Covariance Type:        nonrobust
=====

```

```

=====
                                coef    std err          t      P>|t|
-----
[0.025    0.975]
-----
Intercept                    28.4384    0.068    416.523    0.000
28.305    28.572
cash_type[T.cash]             1.9822    0.082    24.089    0.000
1.821    2.144
coffee_name[T.Americano with Milk]  5.0050    0.041    122.752    0.000
4.925    5.085
coffee_name[T.Cappuccino]      9.8823    0.046    215.775    0.000

```

9.792	9.972				
coffee_name[T.Cocoa]		9.7777	0.057	170.680	0.000
9.665	9.890				
coffee_name[T.Cortado]		0.2849	0.054	5.280	0.000
0.179	0.391				
coffee_name[T.Espresso]		-4.7085	0.071	-66.323	0.000
-4.848	-4.569				
coffee_name[T.Hot Chocolate]		9.9195	0.055	180.433	0.000
9.812	10.027				
coffee_name[T.Latte]		9.8859	0.041	238.916	0.000
9.805	9.967				
time_of_day[T.Morning]		-0.1432	0.030	-4.721	0.000
-0.203	-0.084				
time_of_day[T.Night]		-0.0292	0.031	-0.948	0.343
-0.090	0.031				
weekday[T.Mon]		-0.0092	0.045	-0.207	0.836
-0.097	0.078				
weekday[T.Sat]		0.0452	0.046	0.973	0.330
-0.046	0.136				
weekday[T.Sun]		-0.0026	0.048	-0.055	0.956
-0.097	0.091				
weekday[T.Thu]		-0.0467	0.045	-1.029	0.303
-0.136	0.042				
weekday[T.Tue]		-0.0340	0.044	-0.771	0.441
-0.120	0.052				
weekday[T.Wed]		-0.0339	0.046	-0.743	0.458
-0.123	0.056				
month_name[T.Aug]		-5.4459	0.070	-77.283	0.000
-5.584	-5.308				
month_name[T.Dec]		-2.5121	0.071	-35.384	0.000
-2.651	-2.373				
month_name[T.Feb]		-2.4758	0.065	-37.949	0.000
-2.604	-2.348				
month_name[T.Jan]		-2.5127	0.075	-33.459	0.000
-2.660	-2.366				
month_name[T.Jul]		-4.7861	0.073	-65.955	0.000
-4.928	-4.644				
month_name[T.Jun]		-0.5722	0.073	-7.845	0.000
-0.715	-0.429				
month_name[T.Mar]		-1.3812	0.062	-22.133	0.000
-1.504	-1.259				
month_name[T.May]		-0.6151	0.070	-8.849	0.000
-0.751	-0.479				
month_name[T.Nov]		-2.4990	0.071	-34.998	0.000
-2.639	-2.359				
month_name[T.Oct]		-2.4968	0.065	-38.294	0.000
-2.625	-2.369				
month_name[T.Sep]		-5.1633	0.067	-76.558	0.000

-5.296	-5.031		
--------	--------	--	--

```
=====
```

Omnibus:	1998.605	Durbin-Watson:	0.090
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19703.208
Skew:	2.449	Prob(JB):	0.00
Kurtosis:	13.301	Cond. No.	19.9

```
=====
```

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

AIC full model: 8115.3454454385865

```
[16]: #Stepwise regression (both; based on AIC)
def stepwise_aic(data, response, predictors):
    best_aic = float('inf')
    best_combo = None

    for k in range(1, len(predictors)+1):
        for combo in itertools.combinations(predictors, k):
            formula = "{} ~ {}".format(response, ' + '.join(combo))
            model = smf.ols(formula, data=data).fit()
            if model.aic < best_aic:
                best_aic = model.aic
                best_combo = combo
    return best_combo, best_aic

predictors = ["cash_type", "coffee_name", "time_of_day", "weekday", "month_name"]
best_features, best_aic = stepwise_aic(coffee, "money", predictors)

print("\nBest predictors (AIC):", best_features)
print("Best AIC:", best_aic)

#Reduced model with the selected predictors
reduced_formula = "money ~ " + " + ".join(best_features)
red_mod = smf.ols(reduced_formula, data=coffee).fit()
print(red_mod.summary())
print("AIC reduced model:", red_mod.aic)
```

Best predictors (AIC): ('cash\_type', 'coffee\_name', 'time\_of\_day', 'month\_name')

Best AIC: 8108.529583122794

#### OLS Regression Results

```
=====
```

Dep. Variable:	money	R-squared:	0.978
Model:	OLS	Adj. R-squared:	0.978
Method:	Least Squares	F-statistic:	7560.

Date: Thu, 11 Sep 2025 Prob (F-statistic): 0.00  
Time: 12:15:46 Log-Likelihood: -4032.3  
No. Observations: 3635 AIC: 8109.  
Df Residuals: 3613 BIC: 8245.  
Df Model: 21  
Covariance Type: nonrobust

		coef	std err	t	P> t
[0.025	0.975]				
-----					
Intercept		28.4293	0.062	457.930	0.000
28.308	28.551				
cash_type[T.cash]		1.9856	0.082	24.149	0.000
1.824	2.147				
coffee_name[T.Americano with Milk]		5.0061	0.041	122.991	0.000
4.926	5.086				
coffee_name[T.Cappuccino]		9.8833	0.046	216.129	0.000
9.794	9.973				
coffee_name[T.Cocoa]		9.7783	0.057	171.139	0.000
9.666	9.890				
coffee_name[T.Cortado]		0.2884	0.054	5.358	0.000
0.183	0.394				
coffee_name[T.Espresso]		-4.7128	0.071	-66.491	0.000
-4.852	-4.574				
coffee_name[T.Hot Chocolate]		9.9176	0.055	180.657	0.000
9.810	10.025				
coffee_name[T.Latte]		9.8855	0.041	239.062	0.000
9.804	9.967				
time_of_day[T.Morning]		-0.1460	0.030	-4.834	0.000
-0.205	-0.087				
time_of_day[T.Night]		-0.0350	0.031	-1.141	0.254
-0.095	0.025				
month_name[T.Aug]		-5.4445	0.070	-77.323	0.000
-5.583	-5.306				
month_name[T.Dec]		-2.5106	0.071	-35.384	0.000
-2.650	-2.371				
month_name[T.Feb]		-2.4807	0.065	-38.095	0.000
-2.608	-2.353				
month_name[T.Jan]		-2.5132	0.075	-33.520	0.000
-2.660	-2.366				
month_name[T.Jul]		-4.7892	0.072	-66.097	0.000
-4.931	-4.647				
month_name[T.Jun]		-0.5707	0.073	-7.832	0.000
-0.714	-0.428				
month_name[T.Mar]		-1.3823	0.062	-22.171	0.000
-1.504	-1.260				



month_name[T.May]	-0.6202	0.069	-8.931	0.000
-0.756	-0.484			
month_name[T.Nov]	-2.4946	0.071	-34.982	0.000
-2.634	-2.355			
month_name[T.Oct]	-2.5011	0.065	-38.386	0.000
-2.629	-2.373			
month_name[T.Sep]	-5.1621	0.067	-76.593	0.000
-5.294	-5.030			

=====

Omnibus:	1999.178	Durbin-Watson:	0.090
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19644.994
Skew:	2.451	Prob(JB):	0.00
Kurtosis:	13.280	Cond. No.	19.1

=====

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

AIC reduced model: 8108.529583122794

```
[17]: #To answer Q3
full_formula3 = "money ~ hour_of_day + weekday"
modpredQ3 = smf.ols(full_formula3, data=coffee).fit()
print(modpredQ3.summary())
```

OLS Regression Results

=====

Dep. Variable:	money	R-squared:	0.039
Model:	OLS	Adj. R-squared:	0.037
Method:	Least Squares	F-statistic:	21.22
Date:	Thu, 11 Sep 2025	Prob (F-statistic):	3.26e-28
Time:	12:15:46	Log-Likelihood:	-10876.
No. Observations:	3635	AIC:	2.177e+04
Df Residuals:	3627	BIC:	2.182e+04
Df Model:	7		
Covariance Type:	nonrobust		

=====

==

	coef	std err	t	P> t	[0.025
0.975]					
-----					
--					
Intercept	28.5592	0.334	85.461	0.000	27.904
29.214					
weekday[T.Mon]	0.1936	0.291	0.666	0.505	-0.376
0.763					
weekday[T.Sat]	-0.2369	0.302	-0.784	0.433	-0.829
0.355					

weekday[T.Sun]	0.1212	0.311	0.390	0.697	-0.488
0.731					
weekday[T.Thu]	-0.2269	0.297	-0.765	0.444	-0.808
0.355					
weekday[T.Tue]	0.0566	0.288	0.197	0.844	-0.507
0.621					
weekday[T.Wed]	-0.2458	0.298	-0.826	0.409	-0.829
0.338					
hour_of_day	0.2280	0.019	12.017	0.000	0.191
0.265					

Omnibus:	226.579	Durbin-Watson:	1.488
Prob(Omnibus):	0.000	Jarque-Bera (JB):	194.868
Skew:	-0.494	Prob(JB):	4.84e-43
Kurtosis:	2.443	Cond. No.	109.

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
[2]: import os
      os.environ["PATH"] += os.pathsep + "/Library/TeX/texbin"
```

```
[ ]:
```