Consider yourself working for a global retailer that over the years has added a web-based channel to their physical store locations. Now, after learning more about mobile-led changes in retailing, they are excited about what the mobile ecosystem offers. They are seeking your help as they embark on using mobile as a channel. They want to commission an app development team to deploy a presence on iOS and Android. However, several questions arise about the deployment of the app. Your job is to provide data driven insights to help them navigate this complex landscape.

Specifically, you are tasked with:

1. Using the data, estimate a linear model for the relationship between demand and price. For this you have access to a large volume of app level data (in a file called **hw2_1.csv**), including information about the 'rank' of the app on the app store. Assume Sales = (1/rank)*1,000,000 (don't worry about the details behind this assumption, just make the assumption). Specifically, estimate a univariate regression where the dependent variable is sales and the independent variable is price: $Sales = \beta_0 + \beta_1 * Price$
   a. Report the estimated intercept and the estimated slope coefficient.
   b. Test the following null hypothesis: $\beta_1 = 0$. Use a 5% significance level. Provide an explanation of your answer.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     13.06
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           0.000302
Time:                        16:45:58   Log-Likelihood:            -2.3509e+05
No. Observations:               18624   AIC:                         4.702e+05
Df Residuals:                   18622   BIC:                         4.702e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    2.07e+04    586.506     35.300      0.000    1.96e+04    2.19e+04
price       -469.7850    129.982     -3.614      0.000    -724.562    -215.008
==============================================================================
```

2. Create a dummy/binary variable for region. This variable should have a value of 0 if the region is CN (China) and 1 if the region is US (USA). Estimate a univariate regression of sales on this newly created variable. Provide a screenshot and an **interpretation** of both estimated coefficients. Be specific.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     22.22
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           2.45e-06
Time:                        16:49:11   Log-Likelihood:            -2.3509e+05
No. Observations:               18624   AIC:                         4.702e+05
Df Residuals:                   18622   BIC:                         4.702e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.763e+04    716.421     24.610      0.000    1.62e+04     1.9e+04
us_dummy    5114.4716   1084.992      4.714      0.000    2987.788    7241.155
==============================================================================
```

3. Create another dummy/binary variable for in app advertisements (in_app_ads). This variable should have a value of 1 if the device has in app advertising and a value of 0 if the device does NOT have in app advertising. Estimate a regression of sales on the dummy variable created in part 2 and this newly created dummy variable (all in the same model). Provide a screenshot of the results and provide an interpretation of **all** the coefficients. Be Specific.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     16.61
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           6.23e-08
Time:                        16:50:44   Log-Likelihood:            -2.3508e+05
No. Observations:               18624   AIC:                         4.702e+05
Df Residuals:                   18621   BIC:                         4.702e+05
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept       1.654e+04    788.140     20.988      0.000      1.5e+04    1.81e+04
us_dummy        4964.4506   1085.646      4.573      0.000     2836.486    7092.415
has_in_app_ads  3910.1267   1179.941      3.314      0.001     1597.335    6222.919
==============================================================================
```

4. Estimate a univariate regression of sales on price (similar to part 1) except in this case your model should able to speak in terms of elasticity. By elasticity you want to speak to your management in percentage terms – what is the % change in sales for a % increase in price? (Tip: we do this using log-log-regression models.) Since price can have a value of 0, you will have to adjust the variable. You can do this by adding 1 to each price and then taking the log. Provide a screenshot of the results and provide an **interpretation** for all the coefficients. Be specific. ([https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/) .](https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/)

```
==============================================================================
Dep. Variable:               ln_sales   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     38.35
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           6.03e-10
Time:                        16:51:47   Log-Likelihood:                -26834.
No. Observations:               18624   AIC:                         5.367e+04
Df Residuals:                   18622   BIC:                         5.369e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.9709      0.010    899.535      0.000       8.951       8.990
ln_price      -0.0617      0.010     -6.193      0.000      -0.081      -0.042
==============================================================================
```

5. The app retailer believes that other factors, specifically the filesize, the number of screenshots, and the average rating may also be associated with both sales and price. The retailers want a model that estimates the relationship between price and sales (similar to 4) except they want the impact of the above-mentioned factors to be controlled for. Estimate a model that accomplishes this. Your model should speak in terms of elasticity (same as part 4). Provide screenshots of your results and discuss how this model achieves what the retailers want. Provide an interpretation of **all** the estimated coefficients.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               ln_sales   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     20.54
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           6.59e-17
Time:                        16:53:02   Log-Likelihood:                -26812.
No. Observations:               18624   AIC:                         5.363e+04
Df Residuals:                   18619   BIC:                         5.367e+04
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          8.7695      0.040    219.143      0.000       8.691       8.848
ln_price          -0.0678      0.011     -6.211      0.000      -0.089      -0.046
filesize        8.758e-05   3.49e-05      2.512      0.012    1.92e-05       0.000
num_screenshot    -0.0009      0.003     -0.309      0.757      -0.007       0.005
average_rating     0.0487      0.008      5.920      0.000       0.033       0.065
==============================================================================
```

6. The retailer is also interested in understanding the impact of the in-app purchase option. Specifically, the retailer believes that the relationship between price and sales is different for apps with an in-app purchase option and apps without an in-app purchase option. To do this, estimate the same model that you estimated in part 5 except add an interaction term between price and in app purchase option (dummy variable). Provide the results and an interpretation of **all** the estimated coefficients. Be specific.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               ln_sales   R-squared:                       0.012
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     36.41
Date:                Fri, 03 Dec 2021   Prob (F-statistic):           4.07e-44
Time:                        11:31:43   Log-Likelihood:                -26744.
No. Observations:               18624   AIC:                         5.350e+04
Df Residuals:                   18617   BIC:                         5.356e+04
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                     8.7398      0.040    217.190      0.000       8.661       8.819
ln_price                     -0.0175      0.014     -1.292      0.196      -0.044       0.009
filesize                   3.254e-05   3.71e-05      0.876      0.381   -4.03e-05       0.000
num_screenshot               -0.0077      0.003     -2.498      0.012      -0.014      -0.002
average_rating                0.0394      0.008      4.763      0.000       0.023       0.056
has_in_app_purchase           0.2191      0.020     10.732      0.000       0.179       0.259
has_in_app_purchase:ln_price -0.0681      0.021     -3.217      0.001      -0.110      -0.027
==============================================================================
```

Exercise 1. [2 points] You are interested in examining whether visitors to your website spend, on average, more than 12 minutes browsing the website. While you had not previously kept track of website visitors, you start tracking after deciding that you want this information. In the file exercise_1.csv, you will find the minutes spent browsing for a sample of 24 website visitors. Use this data to statistically evaluate whether, on average, website visitors spend more than 12 minutes browsing on your website. Be specific about your approach (set your alpha-level at 0.05). State the null/alternative hypothesis and your conclusion.

```
stats.ttest_1samp(ex1["times"], 12 , alternative = "greater")
```

```
Ttest_1sampResult(statistic=1.7584069260008834, pvalue=0.045989713634792644)
```

Exercise 2. [8 points]

a. Evaluate whether the new website design visits are statistically more likely to end in a sale than the visits to the original website design. You do not need to include other variables in your model since the customers were randomly assigned. Be specific about your approach (set your alpha-level at 0.05). State the null/alternative hypothesis and your conclusion.

```
old["sale_1_0"].mean()
```

```
0.11530398322851153
```

```
new["sale_1_0"].mean()
```

```
0.25806451612903225
```

```
stats.ttest_ind(new["sale_1_0"], old["sale_1_0"], equal_var = False, alternative="greater")
```

```
Ttest_indResult(statistic=2.9796986747155985, pvalue=0.0017704133759339467)
```

Examine whether there is a statistical difference between the mean of minutes_spent for the subset of consumers that were sent to the new website design and the subset that were sent to the original website design. You do not need to include other variables in your model since the customers were randomly assigned. Be specific about your approach (set you alpha-level at 0.05). State the null/alternative hypothesis and your conclusion.

```
old["minutes_spent"].mean()
```

```
6.465408805031447
```

```
new["minutes_spent"].mean()
```

```
8.978494623655914
```

```
stats.ttest_ind(new["minutes_spent"], old["minutes_spent"], equal_var = False, alternative="two-sided")
```

```
Ttest_indResult(statistic=6.449965845079837, pvalue=2.187142466692432e-09)
```

There is concern that there may have been a programming error regarding the random assignment of consumers. Specifically, it may be that the selection of the 10% of customer traffic that was directed to the newly designed website was not random. Does the data suggest that this concern is legitimate? Even if it was not random, does the data suggest that our conclusions about the new website design observed in part a should change? Be specific about your approach (set you alpha-level at 0.05). State the null/alternative hypothesis and your conclusion.

```
new["member"].mean()
```

```
0.7311827956989247
```

```
old["member"].mean()
```

```
0.26834381551362685
```

```
#Example 2c
stats.ttest_ind(new["member"], old["member"], equal_var = False, alternative="two-sided")
```

```
Ttest_indResult(statistic=9.167498511792347, pvalue=9.237501389379306e-16)
```

```
result = sm.ols(formula="sale_1_0 ~ website_design + member ",
                data=ex2).fit()
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               sale_1_0   R-squared:                       0.210
Model:                            OLS   Adj. R-squared:                  0.207
Method:                 Least Squares   F-statistic:                     75.28
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           1.01e-29
Time:                        17:29:59   Log-Likelihood:                -135.94
No. Observations:                 570   AIC:                             277.9
Df Residuals:                     567   BIC:                             290.9
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          0.0249      0.016      1.547      0.122      -0.007       0.057
website_design    -0.0131      0.037     -0.350      0.726      -0.087       0.060
member             0.3367      0.029     11.569      0.000       0.280       0.394
==============================================================================
```

The company is worried that the introduction of the new features (more product pictures and zoom feature) is having an impact on customers returning the products they purchased. Statistically examine the impact of the new website design on returns.

```
old["return_1_0"].mean()
```

```
0.0649895178197065
```

```
new["return_1_0"].mean()
```

```
0.0967741935483871
```

```
stats.ttest_ind(new["return_1_0"], old["return_1_0"],
                equal_var = False)
```

```
Ttest_indResult(statistic=0.9681829325947716, pvalue=0.334933723658849)
```

```
old[old["sale_1_0"] == 1]["return_1_0"].mean()
```

```
0.5636363636363636
```

```
new[new["sale_1_0"] == 1]["return_1_0"].mean()
```

```
0.375
```

```
stats.ttest_ind(new[new["sale_1_0"] == 1]["return_1_0"], old[old["sale_1_0"] == 1]["return_1_0"],
                equal_var = False)
```

```
Ttest_indResult(statistic=-1.5534782160704874, pvalue=0.12741211164128996)
```

```
result = sm.ols(formula="return_1_0 ~ website_design + sale_1_0 ",
                data=ex2).fit()
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             return_1_0   R-squared:                       0.473
Model:                            OLS   Adj. R-squared:                  0.471
Method:                 Least Squares   F-statistic:                     254.0
Date:                Mon, 29 Nov 2021   Prob (F-statistic):           1.70e-79
Time:                        17:38:38   Log-Likelihood:                 151.45
No. Observations:                 570   AIC:                            -296.9
Df Residuals:                     567   BIC:                            -283.9
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          0.0058      0.009      0.654      0.514      -0.012       0.023
website_design    -0.0415      0.021     -1.944      0.052      -0.083       0.000
sale_1_0           0.5131      0.023     22.490      0.000       0.468       0.558
==============================================================================
```