Final Assignment

Part 1 - Employment

A large university knows that about 70% of the full-time students are employed at least 5 hours per week. The members of the Statistics Department wonder if the same proportion of their students work at least 5 hours per week. They randomly sample 25 majors and find that 15 of the students (60%) work 5 or more hours each week.

Question 1

Describe how you can set up a simulation to estimate the proportion of statistics majors who work 5 or more hours each week based on this sample.

Question 2

A bootstrap distribution with 1000 simulations is shown below. Approximate the bounds of the 95% confidence interval based on this distribution.



Suppose the lower bound of the confidence interval from the previous question is L and the upper bound is U. Which of the following is correct?

a. Between L to U of statistics majors work at least 5 hours per week.

b. 95% of the time the true proportion of statistics majors who work at least 5 hours per week is between L and U.

c. Between L and U of random samples of 25 statistics majors are expected to yield confidence intervals that contain the true proportion of statistics majors who work at least 5 hours per week.

d. 95% of random samples of 25 statistics majors will yield confidence intervals between L and U.

e. None of the above.

Part 2 - Blizzard

In 2020, employees of Blizzard Entertainment circulated a spreadsheet to anonymously share salaries and recent pay increases amidst rising tension in the video game industry over wage disparities and executive compensation. (Source: Blizzard Workers Share Salaries in Revolt Over Pay)

The name of the data frame used for this analysis is **blizzard_salary** and the variables are:

- percent_incr: Raise given in July 2020, as percent increase with values ranging from 1 (1% increase to 21.5 (21.5% increase)
- salary_type: Type of salary, with levels Hourly and Salaried
- annual_salary: Annual salary, in USD, with values ranging from \$50,939 to \$216,856.
- performance_rating: Most recent review performance rating, with levels Poor, Successful, High, and Top. The Poor level is the lowest rating and the Top level is the highest rating.

The top ten rows of blizzard_salary are shown below:

	percent_incr	<pre>salary_type</pre>	annual_salary	$performance_rating$
0	1.0	year	1.0	High
1	1.0	year	1.0	Successful
2	1.0	year	1.0	High
3	1.0	Hourly	33987.2	Successful
4	NaN	Hourly	34798.4	High
5	NaN	Hourly	35360.0	NaN
6	NaN	Hourly	37440.0	NaN
7	0.0	Hourly	37814.4	NaN
8	4.0	Hourly	41100.8	Тор
9	1.2	Hourly	42328.0	NaN

Question 4

Next, you fit a model for predicting raises (percent_incr) from salaries (annual_salary). We'll call this model raise_1_fit. An output of the model is shown below.

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
Intercept	1.869965	0.432035	4.328268	0.000019	1.020397	2.719532
annual_salary	0.000016	0.000005	3.431459	0.000669	0.000007	0.000024

Which of the following is the best interpretation of the slope coefficient?

- a. For every additional \$1,000 of annual salary, the model predicts the raise to be higher, on average, by 1.6%.
- b. For every additional 1,000 of annual salary, the raise goes up by 0.016%.
- c. For every additional \$1,000 of annual salary, the model predicts the raise to be higher, on average, by 0.016%.
- d. For every additional \$1,000 of annual salary, the model predicts the raise to be higher, on average, by 1.87%.

Question 5

You then fit a model for predicting raises (percent_incr) from salaries (annual_salary) and performance ratings (performance_rating). We'll call this model raise_2_fit. Which of the following is definitely true based on the information you have so far?

- a. Intercept of raise_2_fit is higher than intercept of raise_1_fit.
- b. Slope of raise_2_fit is higher than RMSE of raise_1_fit.
- c. Adjusted R^2 of raise_2_fit is higher than adjusted R^2 of raise_1_fit.
- d. R^2 of raise_2_fit is higher R^2 of raise_1_fit.

Question 6

The tidy model output for the raise_2_fit model you fit is shown below.

	Coef.	Std.Err.		t	\
Intercept	2.617865	0.452366	5.7	'87046	
<pre>performance_rating_Poor[T.True]</pre>	-3.499070	1.500230	-2.3	32356	
<pre>performance_rating_Successful[T.True]</pre>	-1.730554	0.361704	-4.7	'84449	
<pre>performance_rating_Top[T.True]</pre>	3.628454	0.730187	4.9	69215	
annual_salary	0.000013	0.000004	3.1	.19855	
	P>	t [O	.025	0.9	75]
Intercept	1.543384e	-08 1.72	8293	3.507	437
<pre>performance_rating_Poor[T.True]</pre>	2.022505e	-02 -6.44	9249	-0.548	892
<pre>performance_rating_Successful[T.True]</pre>	2.493372e	-06 -2.44	1838	-1.019	269
<pre>performance_rating_Top[T.True]</pre>	1.035401e	-06 2.19	2554	5.064	355
annual_salary	1.953363e	-03 0.00	0005	0.000	021

When your teammate sees this model output, they remark "The coefficient for performance_ratingSuccessful is negative, that's weird. I guess it means that people who get successful performance ratings get lower raises." How would you respond to your teammate?

Question 7

Ultimately, your teammate decides they don't like the negative slope coefficients in the model output you created (not that there's anything wrong with negative slope coefficients!), does something else, and comes up with the following model output. Note however that the coefficient is still negative, but this satisfies your friend...

	Coef.	Std.Err.		t	\
Intercept	1.785333	0.509233	3.50)5927	
<pre>performance_rating_Successful[T.True]</pre>	-0.800356	0.439216	-1.82	22238	
<pre>performance_rating_High[T.True]</pre>	1.574196	0.475552	3.3	10248	
<pre>performance_rating_Top[T.True]</pre>	4.569528	0.768132	5.94	18886	
annual_salary	0.000012	0.00004	2.8	54297	
	P>	t [0.	025	0.9	975]
Intercept	5.116857e	-04 0.783	3935	2.786	5732
<pre>performance_rating_Successful[T.True]</pre>	6.923704e	-02 -1.664	1068	0.063	3355
<pre>performance_rating_High[T.True]</pre>	1.025064e	-03 0.639	9030	2.509	9362
<pre>performance_rating_Top[T.True]</pre>	6.327904e	-09 3.059	9009	6.080	047
annual_salary	4.559386e	-03 0.000	004	0.000	020

Unfortunately they didn't write their code in a Quarto document, instead just wrote some code in the Console and then lost track of their work. They remember using the fct_relevel() function and doing something like the following:

```
blizzard_salary['performance_rating'] = pd.Categorical(
    blizzard_salary['performance_rating'],
    categories=[____],
    ordered=True
)
```

What should they put in the blanks to get the same model output as above?

- a. "Poor", "Successful", "High", "Top"
- b. "Successful", "High", "Top"
- c. "Top", "High", "Successful", "Poor"
- d. Poor, Successful, High, Top

Suppose we fit a model to predict percent_incr from annual_salary and salary_type. A tidy output of the model is shown below.

	Coef.	Std.Err.	t	P> t	[0.025	\
Intercept	1.242597	0.570278	2.178932	0.029972	0.121174	
<pre>salary_type_year[T.True]</pre>	0.913343	0.543715	1.679819	0.093844	-0.155845	
annual_salary	0.000014	0.000005	2.958979	0.003287	0.000005	
	0.975]					
Intercept	2.364020					
<pre>salary_type_year[T.True]</pre>	1.982532					
annual_salary	0.000023					

Which of the following visualizations represent this model? Explain your reasoning.

Visualizations of the relationship between percent increase, annual salary, and salary type

- a. Figure 1
- b. Figure 2
- c. Figure 3
- d. Figure 4



Figure 1



Figure 2



Figure 3



Figure 4

Define the term parsimonious model.

Suppose you now fit a model to predict the natural log of percent increase, log(percent incr), from performance rating. The model is called raise_4_fit.

You're provided the following:

```
raise 4 fit coefs['exp estimate'] = np.exp(raise 4 fit coefs['estimate'])
print(raise_4_fit_coefs)
```

	term	estimate	exp_estimate
0	Intercept	-2.088594	0.123861
1	performance_rating_Successful	1.872176	6.502427
2	performance_rating_High	3.110229	22.426176
3	performance_rating_Top	3.854360	47.198390

Based on this, which of the following is true?

a. The model predicts that the percentage increase employees with Successful performance get, on average, is higher by 10.25% compared to the employees with Poor performance rating.

b. The model predicts that the percentage increase employees with Successful performance get, on average, is higher by 6.93% compared to the employees with Poor performance rating.

c. The model predicts that the percentage increase employees with Successful performance get, on average, is higher by a factor of 6.502427 compared to the employees with Poor performance rating.

d. The model predicts that the percentage increase employees with Successful performance get, on average, is higher by a factor of 1.872176 compared to the employees with Poor performance rating.

Question 11

.

Which of the following is the definition of a regression model? Select all that apply.

a.
$$\hat{y} = \beta_0 + \beta_1 X_1$$

b. $y = \beta_0 + \beta_1 X_1$
c. $\hat{y} = \beta_0 + \beta_1 X_1 + \epsilon$
d. $y = \beta_0 + \beta_1 X_1 + \epsilon$

Part 3 - Calculus

Question 12

Compute the derivative $(\frac{d}{dx})$ of the following function (show each step):

$$g(x) = \left(\sin(x^2) + \cos(ax)\right)^k$$

Question 13

Compute the following integral (show each step):

$$\int_{a}^{b} \left(e^{cx} + \frac{1}{x^n} \right) dx$$

Part 4 - Linear algebra

Question 14

Given a vector
$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$
, write down its transpose x^{\top} .

Question 15

Given the following matrix N:

$$N = \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \\ n_{31} & n_{32} \\ n_{41} & n_{42} \end{bmatrix}$$

Write down its transpose, $N^\top.$

Consider the following matrices C and D:

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix}, \quad D = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \end{bmatrix}$$

- 1. What are the dimensions of C?
- 2. What are the dimensions of D?
- 3. For the matrix product CD:
 - 1. Determine if the product is valid, and explain why.
 - 2. If the product is valid, write down the dimensions of the resulting matrix without computing the product.

Question 17

Given the matrices E and F:

$$E = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \\ e_{31} & e_{32} \end{bmatrix}, \quad F = \begin{bmatrix} f_{11} \\ f_{21} \end{bmatrix}$$

- 1. What are the dimensions of E?
- 2. What are the dimensions of F?
- 3. For the matrix product EF:
 - 1. Determine if the product is valid, and explain why.
 - 2. If the product is valid, compute the resulting matrix.

Bonus

Pick a concept we introduced in class so far that you've been struggling with and explain it in your own words.