Question 1: EDA warmup (7 Points)

A.(1) At www.census.gov you can create a population pyramid for any country. These pyramids are back-to-back histograms. This pyramid shows Mexico's 2000 female population and the projection for 2050. Write *a few sentences* in order to summarize to your boss the changes that are forecast. Feel free to use (or not) any of the storytelling examples we discussed in class.
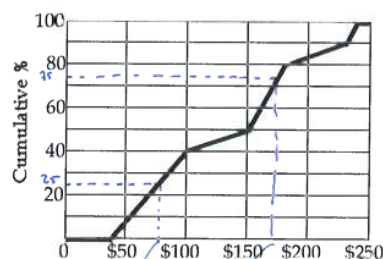

Handwritten notes:
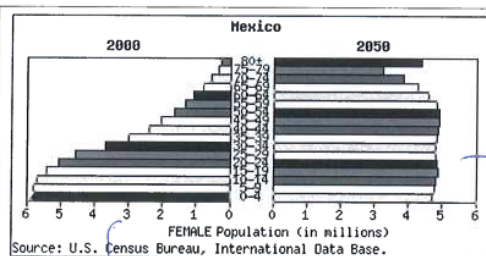
Things to comment:

dramatic change in the population mix

→ balanced/uniform

skewed <30

B. (1)

The weekly expenses for alcohol for a random person are summarized in the graph on the right. Estimate the IQR of these expenses.

i) $50
ii) $75
iii) $100
iv) $150
v) $200

~75   ~175

Explanation (not needed, only if you are not sure):

$$IQR = Q3 - Q1 = 175 - 75 = 100$$

C. (1) Last weekend Maastricht police department ticketed 18 women whose mean speed was 72 km/h, and 30 men going an average of 64 km/h. Overall, what was the mean speed of all the people ticketed?

i)   67 km/h
ii)   68 km/h
iii)   69 km/h
iv)   none of those
v)   it cannot be determined

Explanation (not needed, only if you are not sure):

$$\frac{18 \times 72 + 30 \times 64}{48}$$

D. (4) A publishing company pays its sales staff €600 a week plus a commission of €0.50 per book sold. For example, a salesman who sold 440 books earned 600 + 0.50(440) = €820.

i.(2) The table on the right shows summary statistics for the number of books the sales staff sold last week. Fill in the table to show the statistics for the pay these people earned. Show your calculations below.

Statistic	Books sold	€ Earned
Mean	640	920
Standard Deviation	360	180
IQR	450	225
Maximum	1420	1310

mean: 640 → 600 + 0.50 × 640 = 920

max: 1420 → 600 + 0.50 × 1420 = 1310

!! The sdev & IQR are not affected by the mean

sdev: 360 → 0.50 × 360 = 180

IQR: 450 → 0.50 × 450 = 225

ii.(2) The newest employee had a pretty crazy selling week. Among all the salespeople, her pay corresponded to a *z-score of +1.80*. What was the *z-score* of the number of books she sold? Are you surprised by the result? Explain shortly.

$$Z_{score}^{F} = 1.80 = \frac{x - 920}{180} \Rightarrow \cdots \Rightarrow x = 1244 \in$$

$$books \Rightarrow: \quad 600 + 0.5 \times books = 1244 \Rightarrow books = 1288$$

$$Z_{score}^{B} = \frac{1288 - 640}{360} = \underline{\underline{1.8}}$$

Z are the same since the 2 variables are connected via a linear relationship.

NB: You could answer this directly (no calculations)

Question 2: Traffic Delays (8 Points)

The US administration is trying to fight fake news about traffic jams in USA. A prominent data scientist used data that include information on the *Total Delay per Person* (hours per year spent delayed by traffic), the *Average Arterial Road Speed* (mph), the *Average Highway Road Speed* (mph), and the *Size* of the city (small, medium, large, very large). The regression model based on these variables is presented below. The variables *Small, Large* and *Very Large* are indicators which are 1 for cities of the named size and 0 otherwise.

Dependent variable is: *Total Delay per Person* R-squared = 79.1% R-squared (adjusted) = 77.4%

Variable	Coefficient	SE (Coeff)	t-ratio	P-value
Intercept	139.104	16.69	8.33	<0.0001
Arterial mph	-2.04836	0.6672	-3.07	0.0032
Hiway mph	-1.07347	0.2474	-4.34	<0.0001
Small	-3.5897	2.953	-1.22	0.2287
Large	5.00967	2.104	2.38	0.0203
Very Large	3.41058	3.323	1.06	0.2951

A.(2) Predict the Total Delay for a person with average arterial speed 30 mph, average highway speed 45 mph living in a large city.

B.(1) Explain shortly how we interpret the R-squared in this case.

C.(2) Explain how the coefficients *Small, Large* and *Very Large* account for the size of the city in the model. Why there is no coefficient for Medium? What is the interpretation of the value of coefficient *Large* in **this particular** regression model?

A) total delay $= 139.104 - 2.04836 \times 30 - 1.07347 \times 45 + 5.00967$

$= 34.35672$ hours/year

B) 79.1% of the variance in the delay due to traffic can be explained by the variables used in this model

C). medium is used as a baseline model

• small, large, very large are binary variables that add to the total delay (or reduce it) based on the type of city

• large: In a large city the effect on total delay is ~5 hours, given fixed values for all other variables
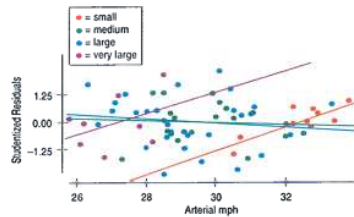
D.(3) Here is a plot of the residuals against the Arterial mph for the model fit above. The plot is colored according to City Size and regression lines are fit for each size.

Considering this display, would you conclude that including the indicator variables for the size of the city accomplished what is needed for the regression model? Explain shortly. *Hint: Think about how we interpret categorical predictors*

In the case of categorical variables what we expect to change the intercept (but the slope to remain the same) see slide 74 of Statistical Larness
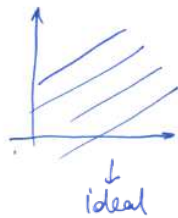
Since we don't see this figure, we assume that it was not that successful

ideal

Question 3: Cats & Dogs (7 points)

Consider the following two documents:

D1:   Dog eat dog. Eat cat too!
D2:   Eat home, it's raining cats and dogs.

A.(1) What would be some standard pre-processing steps applied on these documents and what kind of problems are they expected to solve?

B.(2) Given the pre-processing you did at **A.** what would be the cosine similarity of these documents in a typical information retrieval setup where we only care about raw term frequencies?

C.(2) In a standard cosine-based tf-idf information retrieval system, can a query retrieve a document that does not contain any of the query words? Justify your answer properly.
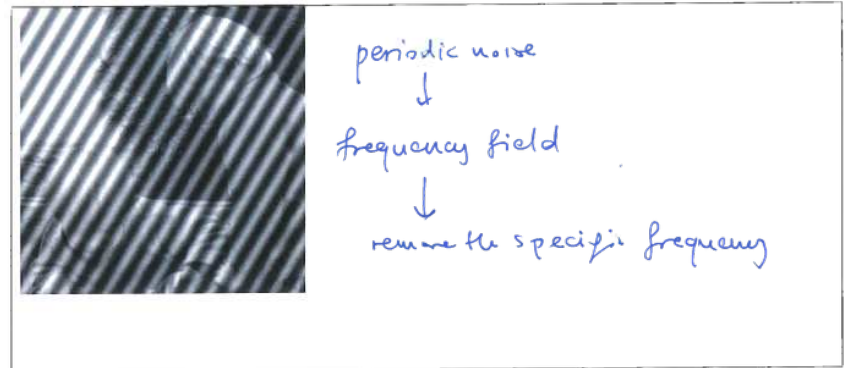

C.(2) You have the following image where we have added extra periodic noise. Explain shortly how you would (most effectively) remove this noise.

periodic noise
↓
frequency field
↓
remove the specific frequency

Question 5: Smart investments (6 points)

Jerry decided to start investing money in specific bonds. In the following plot you can see the daily changes in the return of a high-risk/high-reward bond for 200 consecutive days.

Daily Changes in the Return of an AT&T Bond
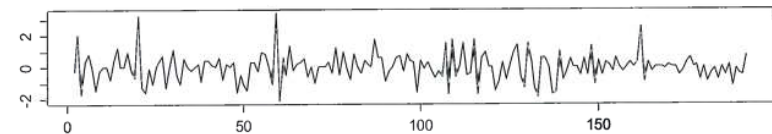
A.(2) Given the autocorrelation plot below, what can you say about the stationarity of the timeseries?
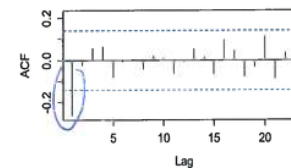
lag1: high value → there is more structure than random

B.(2) Jerry fit an ARIMA(0,0,1) model and the results were the following:

```
Coefficients:
         ma1    intercept
      -0.2865   -0.0247
s.e.   0.0671    0.0426
```

What information is this model using for future prediction? Write clearly the equation of the model and explain all variables involved.

$$\hat{X}_t = -0.0247 - 0.2865\, \hat{e}_t$$

error term

C.(2) In order to find the most critical period of the timeseries (i.e. when important changes happen to the bond), Jerry got the following periodogram (x-axis shows the k-th Fourier frequencies and y-axis the spectral density). Can you identify which are the two most critical periods (in **months**) in this specific dataset?
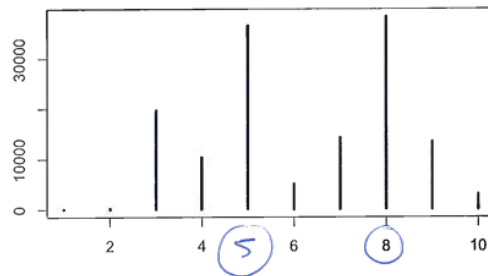
$$K=5 \sim 5/200 = 0.025 \text{ cycles/period} \Rightarrow 40\, days \sim 1.33\, months$$

$$K=8 \rightarrow 8/200 = 0.04 \text{ cycles/period} \rightarrow 25\, days \rightarrow 0.83\, months$$

Question 6: Dimensionality Reduction (7 points)

A.(5) A matrix (Q) contains ratings of 4 users (A,B,C,D) for 4 films (I,II,III,IV) and is decomposed using Singular Value Decomposition (SVD). The data scientist in charge decided to keep just 3 singular values (**13.1, 3.9** and **1.6**) and consequently (by dropping one dimension) the matrices U and V (of SVD) are:

$$U = \begin{array}{c} A \\ B \\ C \\ D \end{array}\begin{bmatrix} -0.48 & -0.16 & -0.06 \\ -0.52 & -0.59 & -0.44 \\ -0.56 & 0.78 & 0.21 \\ -0.44 & -0.12 & 0.87 \end{bmatrix} \quad V = \begin{bmatrix} -0.34 & -0.53 & -0.36 & -0.7 \\ 0.27 & -0.62 & 0.74 & -0.05 \\ 0.86 & -0.18 & -0.47 & -0.04 \end{bmatrix}$$

(I, II, III, IV)

Form (just show, do not compute) how SVD (in the case of using 3 singular values) can reconstruct the original matrix Q. Can you predict the ratings of user **A** for films **I** and **IV**?

$$\underset{\sim}{Q} = U \cdot \Sigma \cdot V = \begin{bmatrix} & U & \end{bmatrix} \times \begin{bmatrix} 13.1 & 0 & 0 \\ 0 & 3.9 & 0 \\ 0 & 0 & 1.6 \end{bmatrix} \times \begin{bmatrix} & V & \end{bmatrix}$$

4×?   3×3   3×4

$$Q_{A-I} = \begin{bmatrix} -0.48 & -0.16 & -0.06 \end{bmatrix} \times \begin{bmatrix} 13.1 & 0 & 0 \\ 0 & 3.9 & 0 \\ 0 & 0 & 1.6 \end{bmatrix} \times \begin{bmatrix} -0.34 \\ 0.27 \\ 0.86 \end{bmatrix} =$$

$$= \begin{bmatrix} -6.288 & -0.624 & -0.096 \end{bmatrix} \times \begin{bmatrix} -0.34 \\ 0.27 \\ 0.86 \end{bmatrix} \cong 1.88$$

Similarly:

$$Q_{A-IV} \cong 4.36058$$

B.(2) Using the example above (and/or other intuitive examples), mention one reason to use SVD versus one reason to use Non-Negative Matrix Factorization (NMF) for the case of movie ratings.

e.g.

SVD: unique decomposition/solution

NMF: non-negative values

Question 7: Regularization magic (7 points)

One bright student tried to fit a Ridge regression model where many of the variables were correlated to each other. The student quickly noticed that Ridge tends to give similar coefficient values to the correlated variables. Let's prove this property with a simple intuitive example.

Suppose that n = 2 (i.e. 2 data points) and p = 2 (i.e. 2 coefficients). Furthermore, we assume that $x_{11} = x_{12}$, $x_{21} = x_{22}$ and that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in (solving the simple linear regression using least squares), is zero: $\hat{\beta}_0 = 0$.

A.(2) Write out the ridge regression optimization problem in this setting (in its simplest form, given the simplifications given above).

$$\text{General Ridge}: \quad \min \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2$$

$$\text{In this setting}: \quad \hat{\beta}_0 = 0, \quad x_{11} = x_{12} = x_1, \quad x_{21} = x_{22} = x_2$$

$$\min \left(y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1 \right)^2 + \left(y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2 \right)^2 + \lambda \cdot \left(\hat{\beta}_1^2 + \hat{\beta}_2^2 \right)$$

$$\underset{\Vert\Omega}{}$$

B.(5) Argue that in this setting (and under the simplifications we did), the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$. *Hint: Attempt solving the optimization problem of the previous question in order to find the β_1 and β_2. Given the simplifications we did, it should be easy to compute.*

The derivative of a function $f(x) = g^2(x)$ is $f'(x) = 2 \ g(x) \ g'(x)$, where g' is the derivative of g.

one possible way:

$$\frac{\partial \Omega}{\partial \hat{\beta}_1} = 2 \cdot (y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1)(-x_1) + 2(y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2)(-x_2) + 2\lambda \hat{\beta}_1 = 0 \Rightarrow$$

$$-y_1 x_1 + \hat{\beta}_2 x_1^2 + x_1^2 \hat{\beta}_1 - x_2 y_2 + \hat{\beta}_2 x_2^2 + x_2^2 \hat{\beta}_1 + \lambda \hat{\beta}_1 = 0 \Rightarrow$$

$$(x_1^2 + x_2^2 + \lambda) \cdot \hat{\beta}_1 = x_1 y_1 + x_2 y_2 - \hat{\beta}_2 (x_1^2 + x_2^2) \Rightarrow$$

$$\hat{\beta}_1 = \frac{x_1 y_1 + x_2 y_2 - \hat{\beta}_2 (x_1^2 + x_2^2)}{\lambda + x_1^2 + x_2^2} \Rightarrow$$

$$X = x_1 = -x_2, \quad y = y_1 = -y_2$$

$$\boxed{\hat{\beta}_1^* = \frac{2(xy - x^2) \cdot \hat{\beta}_1}{\lambda + 2x^2}}$$

If we take $\dfrac{\partial \Omega}{\partial \hat{\beta}_2}$, due to symmetry we will have that

$$\hat{\beta}_2^* = \frac{2(xy - x^2)\hat{\beta}_2}{\lambda + 2x^2}$$

which means that $\underline{\hat{\beta}_1 = \hat{\beta}_2}$

Bonus Question (1 point)

Complete the next verse:

All the leaves are brown / And the sky is grey / I've been for a walk / On a winter's day…

Extra answer sheet.