Warm up with EDA! (13 points)

(7) a. The Board of Examiners of "Super University" collects information about how much time students spend in toilet during exams. Here are the results of some statistics:

mean	7,39 minutes
st.dev.	3,05 minutes
min	3
Q1	3
median	6
Q3	11
max	16

(2) i. Notice that the 1st quartile and minimum number of minutes are the same. Explain how this can be.

(2) ii. Were there any outliers? Explain how you made your decision.

(3) iii. The rules and regulations of Super University declare that students that stay in toilet more than 15 minutes are "suspect for fraud". Would you consider 15 minutes too much? Explain.

i) The lower 25% are 3

e.g. 3, 3, 3, 3, 6, 6, 6, 6, 6, 6, 6, 11, 11, 11, 16 ...

25% = Q1

(ii) I need to compute the fences of boxplot:

$IQR = Q3 - Q1 = 11 - 3 = 8$

Upper Fence: $Q3 + 1.5IQR = 11 + \frac{3}{2} \times 8 = 23$

Lower Fence: $Q1 - 1.5IQR = 3 - \frac{3}{2} \times 8 = -9$

Nothing is < -9 or > 23, so no outliers.

(iii) Compute Z-score:

$$Z = \frac{15 - \mu}{\sigma} = \frac{15 - 7.39}{3.05} = 2.5$$

> 1 kind of rare

o. For the following questions just put in a circle the correct answer. Only if you have doubts, justify your choice in the space provided, otherwise it's perfectly fine to leave it empty.

(2) i) *"If r = –0.4 for the relationship between the time of day and amount of coffee in an office worker's mug, which are true?"*

I. r^2 *(R-squared) = - 16%*
II. *There is a linear relationship between time and amount of coffee.*
III. *In 16% of the cases, amount of coffee in the mug is correctly predicted by time of day.*

A) I
B) II
C) III
D) II and III only
E) none of these

(very short) explanation:
_____
_____
_____
_____

(2) ii) *"The auto insurance industry crashed some test vehicles into a cement barrier at speeds of 5 to 25 mph to investigate the amount of damage to the cars. They found a correlation of r = 0.60 between speed (MPH) and damage ($). If the speed at which a car hit the barrier is 1.5 standard deviations above the mean speed, we expect the damage to be _?_ the mean damage."*

A) equal to
B) 0.36 SD above
C) 0.60 SD above
D) 0.90 SD above
E) 1.5 SD above

(very short) explanation:
_____
_____
_____
_____

(2) iii) *"The correlation between X and Y is r = 0.35. If we double each X value, decrease each Y by 0.20, and interchange the variables (put X on the Y-axis and vice versa), the new correlation"*

A) is 0.35
B) is 0.50
C) is 0.70
D) is 0.90
E) cannot be determined

(very short) explanation:
_____
_____
_____
_____

5 Modeling skiing (16 points)

I have never been skiing and I would love to try it but I am also not fond of crowded places, so I decided to build a model in order to estimate when a small ski resort near Maastricht is crowded. Since I am also very busy, I assigned this problem to some young hard-working students. They collected data from the past two ski seasons, measuring different variables and then they started fighting with each other because everyone claimed to have the best model.

Let's see the solution of a student that came to me having fit a model using the following variables:

Skiers the number of skiers who visit the resort on that day.
Snow the number of inches of snow on the ground.
Temp the high temperature for the day in degrees F.
Weekend an indicator variable, if it's weekend (YES/NO).

Variable	Coefficient	SE (Coeff.)	t-ratio	p-value
Intercept	559.869	76.78	7.29	<0.0001
Snow	1.424	2.70	0.53	0.6019
Temp	-1.604	2.77	-0.58	0.5677
Weekend: YES	147.349	51.86	2.84	0.0086

(2) a. What is the predicted number of skiers for a Saturday with a temperature of 40 F and a snow cover of 25 inches? What would change if it was a Friday?

model is:
$\hat{y} = 559.869 + 1.424 \times Snow - 1.604 \times temp + 147.349 \times weekend$

$\hat{y}_1 = 559.869 + 1.424 \times 25 - 1.604 \times 40 + 147.349 \cong 679$ people

$\hat{y}_2 \cong 531$ people (only the 147.349 changes)

(2) b. Which of the explanatory variables appear to be associated with the number of skiers and which do not? Explain how you reached your conclusion.

- only the weekend seems to have a significant effect

(2) c. Compute a 95% confidence interval (assume for simplicity that *t=2*, i.e. simply compute the intervals as for simple regression) for the slope of the variable **weekend** and explain the meaning of this interval *in the context of this specific problem.*

$147.349 \pm 2 \times 51.86 \rightarrow [44, 251]$

Given fixed values for snow/temperature people on weekends will vary from 44 to 251

(3) d. One of the students is claiming that, given the above results, if we run a single regression where **skiers** remains the dependent variable and **snow** is the only independent variable, we will get a similar result in terms of significance of the coefficient. Do you agree? Why yes/no?

It seems that snow is not significant in the presence of weekend (perhaps it "sucks" the effect of snow)
But if we take the effect of snow by itself the result might be different

(3) e. If you think that the temperature might affect ski attendance differently on weekends than on weekdays, how would you change the regression to test this? Sketch how the regression equation would look like.

Introduce an interaction term
Weekend x temp

$\hat{y} = d_0 + d_1 temp + d_2 snow + d_3 weekend + d_4 weekend \times temp$

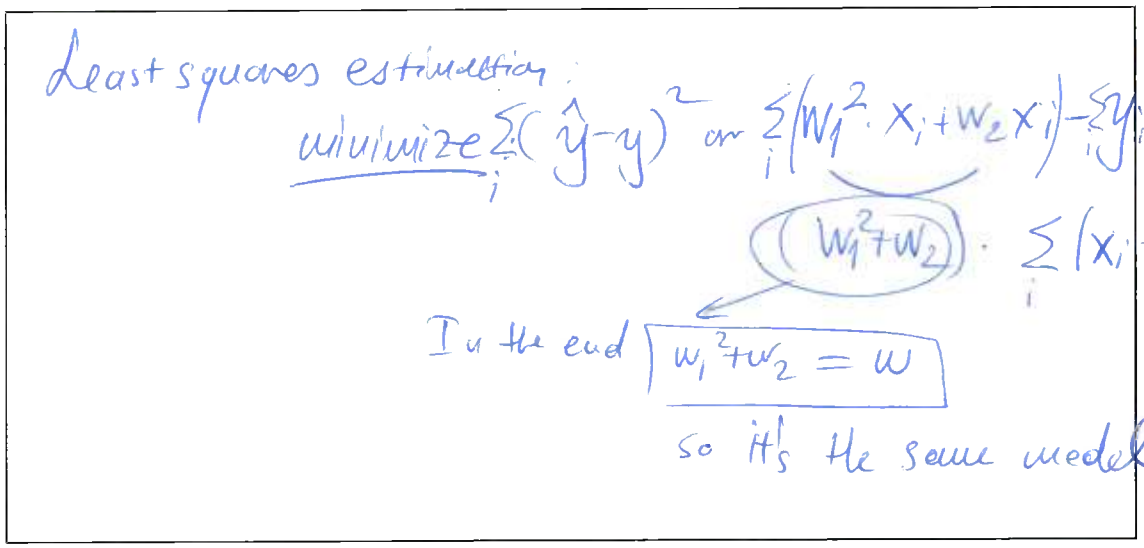
(4) f. Following my advice to fit simple models, two students selected only one variable (**snow**), however they had different opinions on how they should fit the model. Here are the two models they presented:

A: $y = {w_1}^2 x + w_2 x$
B: $y = wx$

Note that model A now uses two parameters (though both multiply with the same input value x). Which of the following is correct?

I) A will perform better than B -most of the times-
II) B will perform better than A -most of the times-
III) They would perform equally well on all cases

Hint: Think of how regression coefficients are estimated.

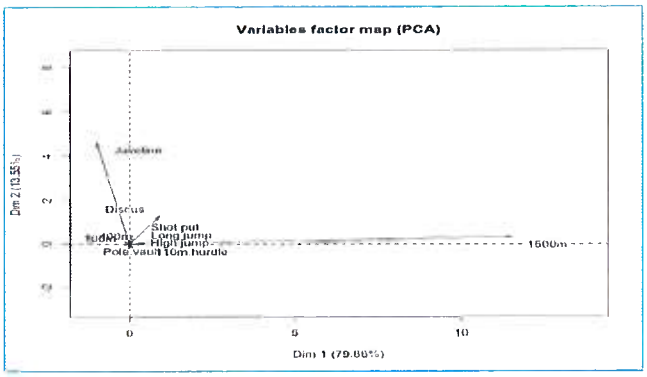


6 Is Decathlon really 10 events? (15 points)

(9) a) Decalthon (from the greek word *"δέκα"* which means *"ten"*) consists of 10 events (namely: 100m, Long Jump, Shot Put, High Jump, 400m, 110m Hurdles, Discus Throw, Pole Vault, Javeline Throw, 1500m). A sports bet company is looking into some data from some athletes in order to reveal which events/sports are most significant.
The chief data scientist of the company (after some googling) decided to run a Principal Component Analysis (PCA), however has no idea on how to interpret the results. The dataset consists of 41 athletes and the performance in 10 sports (see a part of it below).

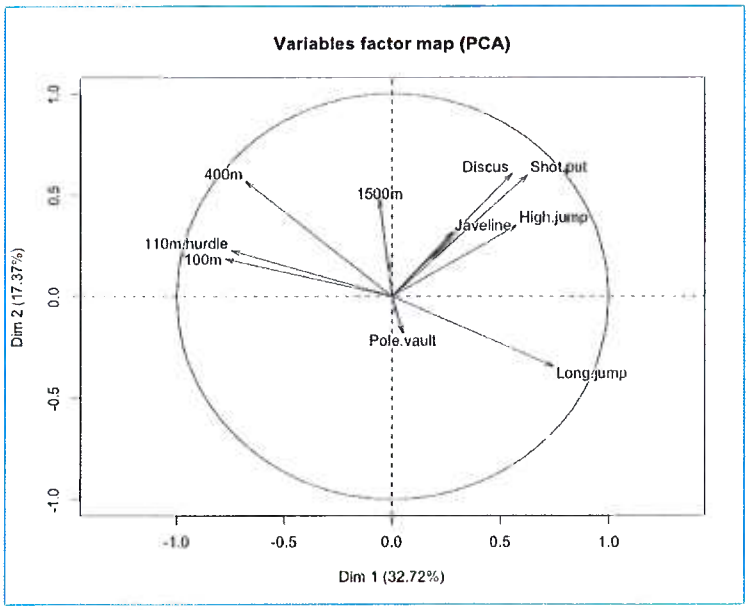
	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.4
WARNERS	11.11	7.60	14.31	1.98	48.68	14.23	41.10	4.92	51.77	278.1

(3) i) The data scientist ran a PCA and by plotting the first 2 PCs got the result on the right. Can you explain the mistake here, why the figure has this form and how you would correct it? Use the data sample above to make assumptions for the data.



- No scaling of the variables
- The 1st dimension is dominated by 1500m which has the largest variance scale
- Rescaling/normality variables

(7) ii) A more reasonable PCA plot using the two first components appears on the right. Answer the following questions.



- Is in this case, is the use of two PCs enough to explain the data? Why yes/no?
- Given the loadings of the figure, what is the 1st and 2nd PC are capturing/ explaining? Can we somehow differentiate between characteristics of different sports? Provide an intuitive description of the above PCA plot.

- First 2 PCs explain around 50%, so it's not enough

- 1st PC ~ jumper
  2nd PC ~ running (100m/150m) } filed & track events

(6) b, We discussed in class that regularization can be applied to any "learning" algorithm in order to avoid overfitting and improve generalization over unseen data. Thinking about Non-Negative Matrix Factorization (NMF), how would you formulate the optimization problem in order to account for regularization? Provide the adjusted optimization equation and the interpretation of parameter shrinking in this case. If we prefer a non-sparse decomposition, would you prefer Ridge (L2) or Lasso (L1) regularization in this case?

normal NMF: minimize $||A - W \cdot H||$

parameters are the actual coefficients

so: minimize: $||A - WH|| + \lambda \cdot ||\mu||$ Lasso

or $\lambda \cdot ||\mu||^2$ Ridge

where $||\mu||$ is a vector with all parameters of W/H matrices

Lasso yields many 0s, so go with Ridge

Bonus question (1 point)

How old is Jerry?

Answer: ____