Problem 1.

(30 points) Imagine that you are interested in the relationship between mood and weather. You ask 4 people to fill out a questionnaire about their mood for 70 consecutive days and also record the maximum temperature of each day. The data for the weather and the mood from the 4 individuals are in the tempmood.csv data set.

(a) Plot the data in a way/s that you find meaningful to examine the relationship between mood and weather. Interpret the graph/s.

Two plots are generated here. The first is a scatter plot showing the regression line between all subjects' mood and the temperature for all days ($n = 4$). We find $r = -.24$, suggesting that only 6% of the variance in mood can be explained by temperature. This plot is useful to show us the overall structure of the data.

A second plot was made to investigate the correlations between individual subject's mood and the weather. Each subject shows a different correlation, $r_{T1} = .3, r_{T2} = 0, r_{T3} = -.3, r_{T4} = -.7$. Correlations between all pairs of subjects are negligible.

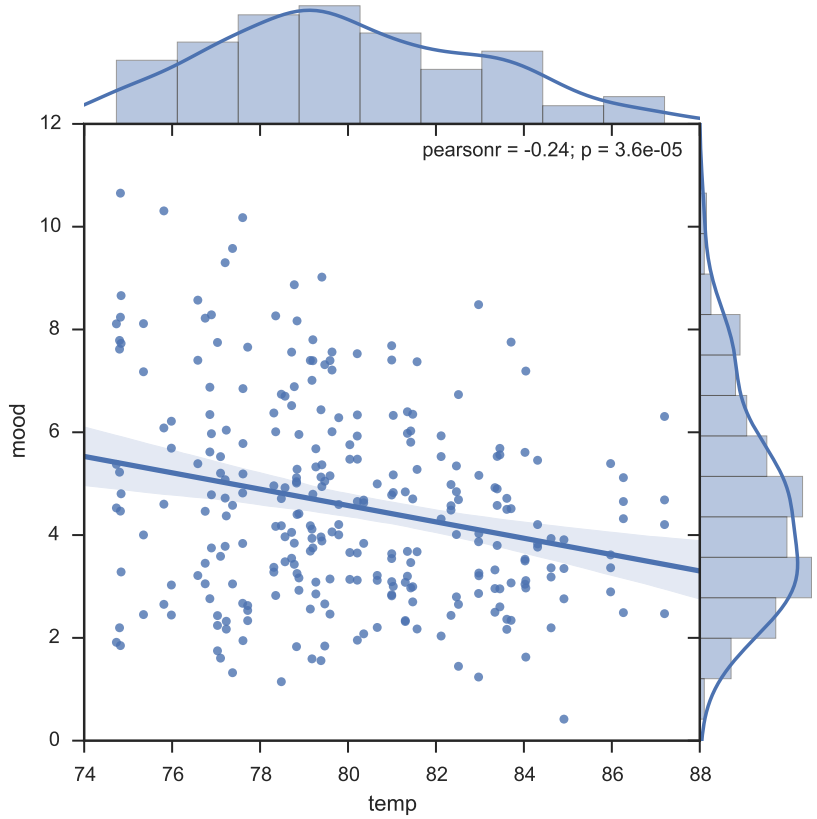

Figure 1: Scatter plot of each subject's mood vs temperature of for each day. Regression line shows a negative correlation between mood and temperature ($r = -0.24$).

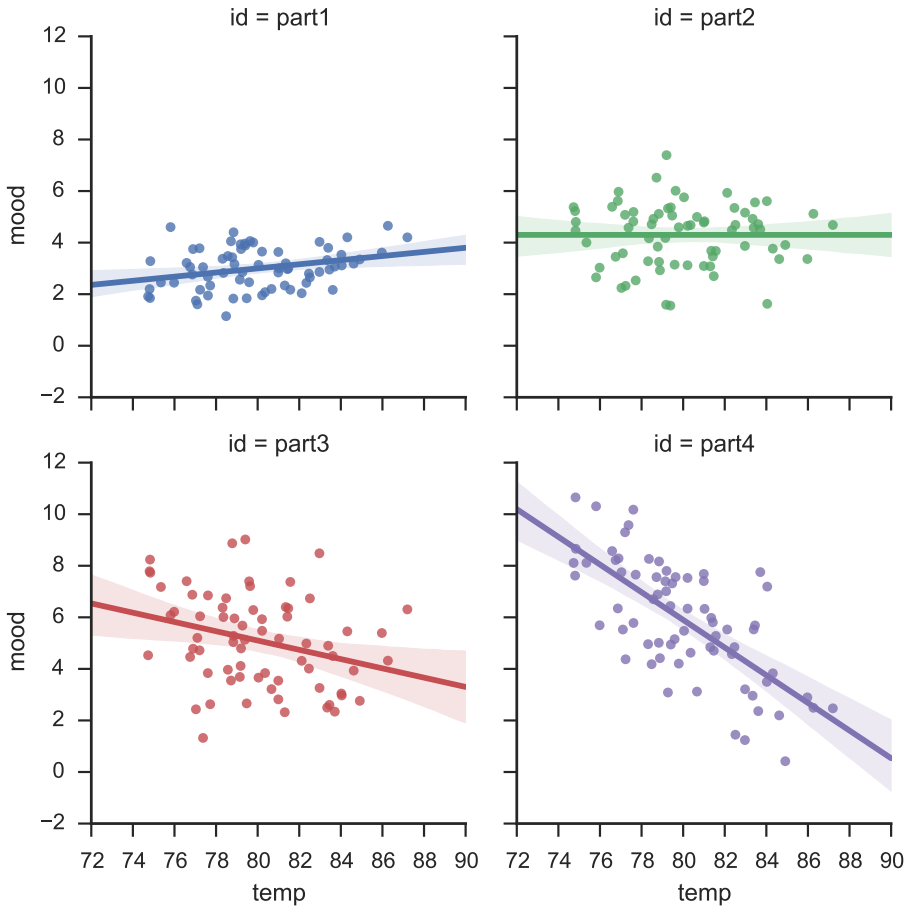

Figure 2: Investigation of individual participant's relationships between mood and temperature. Correlations vary between subjects: $r_{T1} = .3, r_{T2} = 0, r_{T3} = -.3, r_{T4} = -.7$, and correlations between all pairs of subjects are negligible.

(b) Compute means and standard deviations for the data.

```
stats = pd.DataFrame.from_dict({'mean':df.mean(), 'std':df.std()})
print(stats)
```

	mean	std
temp	80.0	3.0
part1	3.0	0.8
part2	4.3	1.2
part3	5.1	1.8
part4	5.9	2.3

(c) Estimate the covariance between weather and mood as well as the sum of cross-products. What does each of these indices indicate about the relation between weather and mood?

```
print(df.cov())

               temp         part1         part2         part3         part4
temp   9.000000e+00  7.200000e-01  2.569603e-15 -1.620000e+00 -4.830000e+00
part1  7.200000e-01  6.400000e-01  1.255035e-15 -8.753063e-16 -4.182645e-15
part2  2.569603e-15  1.255035e-15  1.440000e+00  2.516506e-15  3.946923e-15
part3 -1.620000e+00 -8.753063e-16  2.516506e-15  3.240000e+00 -1.042644e-15
part4 -4.830000e+00 -4.182645e-15  3.946923e-15 -1.042644e-15  5.290000e+00

print(df.cov().sum())

temp     3.27
part1    1.36
part2    1.44
part3    1.62
part4    0.46
```

(d) What is the correlation between weather and mood across all participants? What does the correlation indicate about the relationship between weather and mood?

```
print(df.corr())

               temp         part1         part2         part3         part4
temp   1.000000e+00  3.000000e-01  6.954537e-16 -3.000000e-01 -7.000000e-01
part1  3.000000e-01  1.000000e+00  1.290567e-15 -6.011473e-16 -2.260934e-15
part2  6.954537e-16  1.290567e-15  1.000000e+00  1.165049e-15  1.440538e-15
part3 -3.000000e-01 -6.011473e-16  1.165049e-15  1.000000e+00 -2.837159e-16
part4 -7.000000e-01 -2.260934e-15  1.440538e-15 -2.837159e-16  1.000000e+00
```

(e) Estimate the regression equation of the line that best represents the relationship between weather and mood for all individuals as a single group.

```
import statsmodels.api as sm
def fit_line(x, y):
    """Return slope, intercept of best fit line."""
    X = sm.add_constant(x)
    model = sm.OLS(y, X)
    fit = model.fit()
    return fit.params[1], fit.params[0]

m, b = fit_line(d.temp, d.mood)
print("Slope: {:.2f} Intercept: {:.2f}".format(m, b))

Slope: -0.16 Intercept: 17.31
```

(f) Estimate the regression equation of the line that best represents the relationship between weather and mood for each individual.

```
for i in range(1, 5):
   print("Subject: {:d}, Slope: {:+.2f} Intercept: {:+.2f}"
         .format(i, *fit_line(df.temp, df['part' + str(i)])))

```

```
Subject: 1, Slope: +0.08 Intercept: -3.40
Subject: 2, Slope: +0.00 Intercept: +4.30
Subject: 3, Slope: -0.18 Intercept: +19.50
Subject: 4, Slope: -0.54 Intercept: +48.83
```

(g) Test whether the relation between temperature and mood is significantly different between persons using Fisher's z transformations and z-test.

(h) What can you say about differences in the relationship between weather and mood across individuals?

(h) You submit the result of all these analyses for publication but the editor rejects the manuscript on the basis of: (i) a lack of power to examine your research questions, and (ii) the fact that there are only 4 individuals in your data and, thus — they claim — you cannot generalize to the population. Nevertheless, you are convinced — or just have a hunch — that there might be something valuable here and write back arguing that the data and analyses are worth disseminating. What would you say to support your argument?

Problem 2.

(10 points) The following matrices RV and CV are a correlation and a covariance matrix, respectively, of variables X_1, X_2, X_3, X_4, and X_5. Using the information provided in the matrices, fill in the gray boxes with the appropriate values. Make a note of any anomalies you notice (if any).

		X_1	X_2	X_3	X_4	X_5
Covariance	X_1	4	-	-	-	-
	X_2	-0.5	0.25	-	-	-
	X_3	1.8	1.125	9	-	-
	X_4	-1.08	-0.135	-2.43	7.29	-
	X_5	17.28	1.35	6.48	4.374	29.16
Correlation	X_1	1.0	-	-	-	-
	X_2	-0.5	1.0	-	-	-
	X_3	0.3	0.25	1.0	-	-
	X_4	-0.2	-0.1	-0.3	1.0	-
	X_5	1.6	0.5	0.4	0.3	1.0

There is one anomaly, the correlation $r_{15} = 1.6 > 1$ is impossible. This stems from c_{15} being artificially too large.

$$s_2^2 = 0.25 \implies s_2 = 0.5$$
$$s_4^2 = 7.29 \implies s_4 = 2.7$$

$$r_{11} = \frac{c_{11}}{s_1 s_1} \implies c_{11} = r_{11} s_1 s_1$$
$$r_{13} = \frac{c_{13}}{s_1 s_3}$$
$$r_{14} = \frac{c_{14}}{s_1 s_4}$$
$$r_{15} = \frac{c_{15}}{s_1 s_5}$$
$$r_{23} = \frac{c_{23}}{s_2 s_3} \implies c_{23} = r_{23} s_2 s_3$$
$$r_{24} = \frac{c_{24}}{s_2 s_4} = \frac{-0.135}{(0.5)(2.7)} = -0.1$$
$$r_{25} = \frac{c_{25}}{s_2 s_5} \implies c_{25} = r_{25} s_2 s_5$$
$$r_{33} = \frac{c_{33}}{s_3 s_3} \implies c_{33} = r_{33} s_3 s_3$$
$$r_{35} = \frac{c_{35}}{s_3 s_5}$$
$$r_{55} = \frac{c_{55}}{s_5 s_5} \implies c_{55} = r_{55} s_5 s_5$$

$$\implies$$

$$r_{12} = \frac{c_{12}}{s_1 s_2} \implies s_1 = \frac{c_{12}}{r_{12} s_2} = \frac{-0.5}{(-0.5)(0.5)} = 2$$
$$r_{34} = \frac{c_{34}}{s_3 s_4} \implies s_3 = \frac{c_{34}}{r_{34} s_4} = \frac{-2.43}{(-0.3)(2.7)} = 3$$
$$r_{45} = \frac{c_{45}}{s_4 s_5} \implies s_5 = \frac{c_{45}}{r_{45} s_4} = \frac{4.374}{(0.3)(2.7)} = 5.4$$

$$c_{11} = s_1^2 = 4$$
$$r_{13} = \frac{1.8}{(2)(3)} = 0.3$$
$$r_{14} = \frac{-1.08}{(2)(2.7)} = -0.2$$
$$r_{15} = \frac{17.28}{(2)(5.4)} = 1.6$$
$$c_{23} = (0.25)(0.5)(9) = 1.125$$
$$c_{25} = (0.5)(0.5)(5.4) = 1.35$$
$$c_{33} = s_3^2 = 9$$
$$r_{35} = \frac{6.48}{(3)(5.4)} = 0.4$$
$$c_{55} = s_5^2 = 29.16$$

Problem 3.

(20 points) Researchers were interested in the role of extracurricular activities (sports: 0 = other extracurricular activities and 1 = participation in sports) and biological sex (female: 0 = male, 1 = female) on standard normal adolescent perceptions of social acceptance (PSA). The data can be found in the socialacceptance.csv file. Determine whether factors of extracurricular activity type and biological sex are associated with adolescent PSA.

a) State the type of design of the study.

This is a two-factor study on standard normal adolescent perceptions of social acceptance (PSA). The first factor is involvement with extracurricular activities (sports or 'other'), the second factor is biological sex (female or male).

b) Thoroughly analyze these data for main effects and interactions. Write a report (no longer than a page) in which you report your findings as you would in a journal article (i.e., text, table, and figure).

200 students ($N_{male} = 102, N_{female} = 98$) were investigated to determine whether factors of extracurricular activity type and biological sex are associated with adolescent PSA. Adolescent PSA was subjected to a two-way analysis of variance having two levels of biological sex (female, male) and two levels of extracurricular activity (sports, 'other'). All effects were statistically significant at the $\alpha = .05$ significance level.

The main effect of biological sex yielded an F ratio of $F(1, 196) = 4.27, p < .04$, indicating that the mean adolescent PSA was significantly greater for females ($M = 0.15, SD = 1.09$) than for males ($M = -0.14, SD = 0.88$). The main effect of extracurricular activty type yielded an F ratio of $F(1, 196) = 30.33, p < .01$, indicating that the mean adolescent PSA was significantly higher for sports ($M = 0.36, SD = 0.97$) than for other extracurriculars ($M = -0.36, SD = 0.90$). However, the interaction effect was also significant, $F(1, 196) = 6.15, p = .01$.

An analysis of simple effects showed that the effect of biological sex was significant for sports, $F(1, 99) = 10.1, p < 0.01$, but not for other extracurriculars, $F(1, 97) = 0.1, p = 0.75$. Therefore, there is no evidence that biological sex effects PSA for students participating in other extracurriculars. For females, sports extracurriculars showed a larger adolescent PSA ($M = 0.65, SD = 0.99$) than for males ($M = 0.06, SD = 0.86$). The descriptive statistics for these analyses are presented in Tables 1 and 2.

Source	SS	df	F	$PR(>F)$
Female	3.58	1	4.27	3.99-02
Sports	25.41	1	30.33	1.13-07
Interaction	5.15	1	6.14	1.39-02
Residual	164.24	196		

Table 1: Factorial ANOVA Results for Adolescent PSA Study

Source	SS	df	F	$PR(>F)$
Extracurricular				
Sports	8.7	1	10.1	< 0.01
Other	0.1	1	0.1	0.75

Table 2: Simple Effects Analysis for Adolescent PSA Study

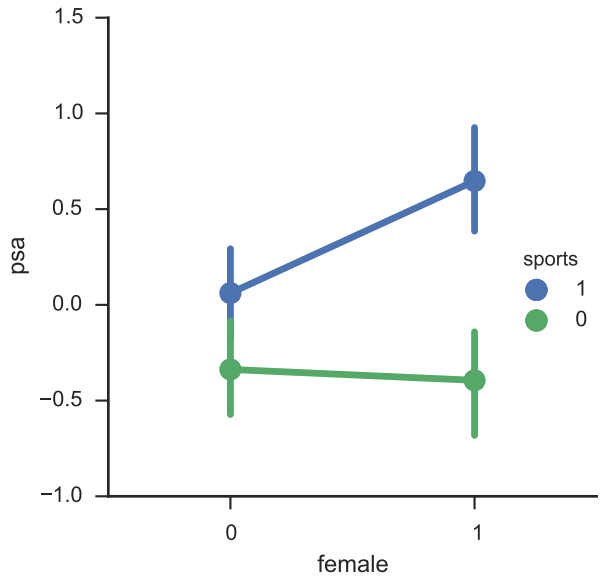

Figure 3: Effects of biological sex ($female = 1, male = 0$) and extracurricular activity ($sports = 1, other = 0$). There is no effect of gender on other extracurriculars, but there is a significant effect of gender on sports extracurriculars. Error bars represent 95% CI.

Problem 4.

(10 points) Explain why r must be between -1 and +1. Please, do not use more than 1 or 2 paragraphs. You can append calculations, if you need them.

r is defined as

$$r = \frac{\text{cov}_{XY}}{s_X s_Y}.$$

As a corrolary of the Cauchy-Schwarz inequality,

$$\begin{aligned} |\text{cov}_{XY}|^2 &= |\mathbb{E}[(X-\mu)(Y-\nu)]|^2 \\ &= |\langle X-\mu, Y-\nu\rangle|^2 \\ &\leq |\langle X-\mu, X-\mu\rangle\langle X-\mu, X-\mu\rangle|^2 \\ &= \mathbb{E}[(X-\mu)^2]\mathbb{E}[(Y-\nu)^2] \\ &= s_X^2 s_Y^2, \end{aligned}$$

where $\mu = E(X), \nu = E(Y)$. From above, we then have $\text{cov}_{XY} = \pm s_X s_Y$. Plugging this into our equation for r yields the solution,

$$\begin{aligned} r &= \frac{\text{cov}_{XY}}{s_X s_Y} \\ &= \frac{\pm s_X s_Y}{s_X s_Y} \\ &= \pm 1 \end{aligned}$$

Problem 5.

(20 points) Say you run a simple regression with predictor variable X_1 and outcome variable Y. You fit the following model:

$$Y = B_0 + B_1 X_1$$

a) What is the interpretation of B_0? What is the interpretation of B_1?

b) When will B_1 be equal to the correlation between Y and X_1? Why?

After running the analysis you remember a covariate that you believe is related to Y, but is not substantively of interest. c) What are the benefits of including the covariate in the model? Include two benefits and explain them in detail.

Finally, you include the covariate within the analysis and fit the following model:

$$Y = B_0 + B_1 X_1 + B_2 X_2$$

d) What is the interpretation of each of the coefficients in this model?

e) When will B_1 be equal to the correlation between Y and X_1? When will B_2 be equal to the correlation between Y and X_2? Why?

Problem 6.

(10 points) Suppose you are hired to serve as a statistical consultant. In each of the following cases, what advice (if any) would you give to your client concerning the procedures and/or conclusions he or she has drawn, or about the kind of statistical techniques most suitable? Be sure to briefly explain the reasoning underlying your advice.

(a) A researcher studies the effects of education (HS or less, Some College, 4 Year College Degree, Graduate/Professional Degree) on income by randomly calling 5,000 participants in the United States. At a presentation of his results several colleagues suggest that effects of education on income may not be robust when considering other predictors such as work experience, time with their current employer, age, personal investments. What sort of analysis did the researcher conduct, and how can the researcher address these criticisms of his research?

(b) A researcher is interested in predicting the mental health of college students based on their reported level of stress. What kind of sample should she collect and what statistical technique should she use to achieve this goal?

(c) A researcher collected data from undergraduate and graduate students at universities across the country in a study of the relation between age (Range of 18 — 46 years with a Mean = 23.5) and openness. There was a significant, negative relation between age and openness ($r(2, 998) = -0.13, p < .05$). The researcher cited this finding as evidence for why elderly individuals (age 60 years and upward) have difficulty learning about novel technology and ideological shifts; they're openness has declined substantially over their lives. Is this a reasonable conclusion? Why or why not?

(d) A researcher studied a group of 100 students by having them complete a survey once a quarter, every quarter, for two years via an online survey form. The survey consisted of several items meant to measure anxiety, self-competence, and academic performance. What methods of analysis would be applicable to this type of data? How do you justify your recommendations?

(e) A researcher received a small grant to conduct a study and is debating on how to spend the money. She is thinking that she can give a test to 300 individuals on one occasion, give a test to one individual on 300 occasions, give a test to 30 individuals on 10 occasions, give a test to 10 individuals on 30 occasions, or any combination of the above. Which of these data collection methods should she use?

Problem 7. Extra Credit

(10 points) Explain what it means to say that a correlation is a covariance expressed in z-scores? Derive numerically the formula for a correlation based on the formula from a covariance (and describe the steps in your own words).