Problem 1.

In your own words describe how RM ANOVA can be used to test for differences between groups, differences between repeated observations, and differences in observations as a function of group membership. These tests have specific names we discussed in lecture.

Problem 2.

Why do we say that in RM ANOVA time is treated as a categorical variable?

Problem 3.

What is the difference between polynomial contrasts and comparisons of means between different observation periods?

Problem 4.

Researchers were interested in how confidence in students fluctuates from the end of sophomore year through the end of senior year in college. Further, researchers were interested in whether students attended a private (i.e., Harvard, Stanford, and Yale) or public (i.e., UMass Boston, San Francisco State University, and Southern Conneticut State University) university would influence end-of-year student confidence ratings, and possible trajectories.

Use the data set confidence.csv to answer the following questions. For your information, t1-t3 indicate time of observation 1-3; i.e., end of sophomore year, end of junior year, and end of senior year, respectively. The variable public indicates whether the student attended a public (public = 1) or private (public = 0) university.

1. Convert the wide format data set to a long format data set. Remember that you will need to create an ID variable, and an Observation/Time variable. Show syntax and the header and footer of the long format data set.

```
import pandas as pd

df = pd.read_csv('confidence.csv')

# Convert to long format
r = pd.melt(df.reset_index(),
            value_vars=['t1', 't2', 't3'],
            id_vars=['index', 'public'],
            var_name='Time', value_name='Observation')

# Some basic sorting
r = r.sort(['index', 'Time']).reset_index(drop=True)

# Rename first column to ID
cols = r.columns.tolist()
cols[0] = 'ID'
r.columns = cols

# Print head and tail of new dataframe
print(r.head())
print(r.tail())
```

OUTPUT

```
   ID  public Time  Observation
0   0       0   t1            4
1   0       0   t2            6
2   0       0   t3            6
3   1       0   t1            3
4   1       0   t2            5

     ID  public Time  Observation
295  98       1   t2            4
296  98       1   t3            4
297  99       1   t1            4
298  99       1   t2            5
299  99       1   t3            4
```

2. Test whether public and private universities differed in their confidence scores. Conduct any pairwise comparison necessary. Report your conclusions.

```
print(anova_lm(ols("Observation ~ C(public)", df).fit(), typ=2))
```

OUTPUT

```
               sum_sq   df          F        PR(>F)
C(public)   48.803333    1  32.419994  2.975196e-08
Residual   448.593333  298        NaN           NaN
```

A one-way between subjects ANOVA was conducted to compare the effect of different types of universities on student confidence for students at public and private universities. There was a significant effect of amount of sugar on words remembered at the $p < .001$ level for the two conditions $[F(1, 298) = 32.4, p = 2.98e-08]$. Comparisons of the mean indicated that the mean confidence scores for the private university students $(M = 3.77, SD = 1.30)$ was significantly different than the public university students $(M = 2.96, SD = 1.15)$.

3. Test whether confidence differs over time. Do this by comparing all observations to each other. Also do this by testing the maximum allowable number of polynomial contrasts. Report your conclusions for both tests.

```
from patsy import dmatrix

# This is equivalent to R's contr.poly
p = dmatrix("C(df.Time, Poly())", df)
poly = pd.DataFrame(p, columns=['Intercept', 'Linear', 'Quadratic'])
df = pd.concat((df, poly), axis=1)
print(anova_lm(ols("Observation ~ Time", data=df).fit()))
print(anova_lm(ols("Observation ~ Linear + Quadratic", data=df).fit(), typ=2))
```

OUTPUT

```
           df      sum_sq   mean_sq         F    PR(>F)
Time        1    7.605000  7.605000  4.627049  0.032275
Residual  298  489.791667  1.643596       NaN       NaN

               sum_sq   df         F    PR(>F)
Linear       7.605000    1  4.666801  0.031551
Quadratic    5.801667    1  3.560187  0.060156
Residual   483.990000  297       NaN       NaN
```

4. Test for a possible interaction between variables tested in question 2 (university type) and 3 (mean differences between observations and significant trends in confidence). Conduct any follow-up analyses necessary if there is a significant interactions. Report your conclusions for all tests.

```
print(anova_lm(ols("Observation ~ C(Time)*C(public)", data=df).fit(), typ=2))

# Test for simple effects
print(anova_lm(ols("Observation ~ C(Time)", data=df.query('public == 0')).fit(), typ=2))
print(anova_lm(ols("Observation ~ C(Time)", data=df.query('public == 1')).fit(), typ=2))
```

```python
print(anova_lm(ols("Observation ~ C(public)", data=df.query('Time == 1')).fit(), typ=2))
print(anova_lm(ols("Observation ~ C(public)", data=df.query('Time == 2')).fit(), typ=2))
print(anova_lm(ols("Observation ~ C(public)", data=df.query('Time == 3')).fit(), typ=2))
```

OUTPUT

```
                       sum_sq   df          F        PR(>F)
C(Time)             13.406667    2   4.995134  7.356729e-03
C(public)           48.803333    1  36.366858  4.897455e-09
C(Time):C(public)   40.646667    2  15.144371  5.496327e-07
Residual           394.540000  294        NaN           NaN

              sum_sq   df          F    PR(>F)
C(Time)    37.213333    2  12.685187  0.000008
Residual  215.620000  147        NaN       NaN

          sum_sq   df        F    PR(>F)
C(Time)    16.84    2  6.91784  0.001345
Residual  178.92  147      NaN       NaN

           sum_sq  df         F    PR(>F)
C(public)    0.81   1  1.107113  0.295298
Residual    71.70  98       NaN       NaN

           sum_sq  df         F    PR(>F)
C(public)    4.00   1  2.911468  0.091119
Residual   134.64  98       NaN       NaN

           sum_sq  df          F        PR(>F)
C(public)   84.64   1  44.073964  1.766331e-09
Residual   188.20  98        NaN           NaN
```

Source	SS	df	F	$PR(> F)$
Time	13.40	2	5.00	$< .001$
Public	48.80	1	36.37	$< .001$
Interaction	40.64	2	15.14	$< .001$
Residual	394.54	294		

Table 1: Factorial ANOVA Results

5. Create a plot of the data that represents the trajectory for public and private students. Make sure the trajectories have some indication of variability around the average trajectory. Sufficiently label and describe your figure.

Source	SS	df	F	$PR(> F)$
Public				
Private	37.21	2	12.7	$< .001$
Public	16.84	2	6.9	.001
Time				
1	0.81	1	1.1	.001
2	4.00	1	2.9	$< .295$
3	84.64	1	44.1	$< .001$

Table 2: Results of Simple Effects Analysis