Overall comments:

Very good job on this project! I was very pleased with your work, and it was clear that you put a lot of effort into this. The grading sheets that I handed back show the distribution of points, and comments on your memos refer to specific numbered comments, which appear below. Even if I don’t ask you to look at specific comments, you should read through this entire list before finishing Part II.

In Part II, I will tend to grade harder on the various factors (and Part II will be worth 60% of the total PIT grade) because ideally you will have taken advantage of the feedback you’re getting from Part I.

Don’t forget: for Part II, you don’t need to report on general characteristics of the youth (e.g., gender, race/ethnicity, etc.), but you do need to show and discuss descriptive statistics for all the variables you use in your analyses. Furthermore, don’t forget to comment on any missing data (including the reasons it might be missing and whether this is likely a concern for interpreting your findings).

1. **Use of tables for descriptive statistics**: It is not necessary (nor desirable) to create a separate table for each variable. Instead, it is often better to create a single table (usually attached at the end of the memo or paper) that includes descriptive statistics for all the variables you are using in the analysis. You can refer to this table in the memo when you are discussing the descriptive statistics.

   Furthermore, you can include a fair amount of detail in the table (e.g., means, median, mode, std dev, max, min, frequencies when appropriate, but only highlight general trends and particular statistics in the memo.

   One last point – many of you included a separate frequency table for each variable. This probably isn’t necessary: measures of central tendency and dispersion are often more concise to convey information about the data. You may, in some cases, want to show the entire frequency distribution (through a table or a figure), but don’t forget about your friends the summary stats.

2. **Clearly label tables, charts, and graphs, and refer to them at appropriate points in the memo**. Just saying “see attached table” or “see attached graph” is not sufficient: if you have multiple tables, charts, and graphs, the reader will have to hunt around for the one that seems most relevant. You should label and number tables, charts, figures and/or graphs, give them names, and refer to them in the memo text when you want the reader to look at them. Tables and Figures are usually numbered separately.
Another note: tables and figures should always be placed physically after you refer to them in the memo, never before.

And another: you should never include a table or figure that you do not refer to in the memo. In other words, every table or figure that you include must be referred to in the memo at some point.

3. **Use of decimal points.** When reporting percentages, it is often not necessary in your memo to report out to 2 decimal points. E.g., it is sufficient to say “51 percent of the sample is male.” When reporting, e.g., number of persons in youths’ families, it is sufficient to say “On average, youth have 4.5 family members, with a standard deviation of 1.5 family members.”

4. **Provide descriptive statistics for all the variables you analyze.** In providing a general description of youth in the NLSY97, I asked you to select two additional variables, and further asked that the variables you choose not simply be the ones that you used in the analyses for part 2. However, this did not mean that you were off the hook for providing descriptive statistics for family income, cigarette smoking, and BMI (see page 3 of the assignment).

5. **The word “data” is a plural word.** Thus, for example, you would say “data are…,” “these data…,” “data indicate…,” NOT “data is…,” “this data…,” “data indicates….”

6. **Extent of missing data for family income, health insurance status, AFDC receipt, etc.** Most often these variables came from the youth’s parent’s portion of the survey, not directly from the youth. Of the 8,984 youth respondents, about 1,000 parents refused to provide answers to any of the parent survey questions. Thus, it’s often difficult to infer much about why they didn’t answer any particular question in the parent survey.

7. **Numbers and Percentages.** See this link for guidelines of when to spell out numbers and when to use actual numbers, as well as many other useful tidbits:
http://memory.loc.gov/ammem/ndlpedit/handbook/numberdate.html

When reporting a percent, many publications ask that the % sign be spelled out – e.g., “51 percent” instead of “51%.”

For either numbers or percentages, if the value is less than one, you should put in a leading zero to avoid confusion about whether a sentence just ended: e.g., “0.71” instead of “.71”

8. **Further interpret strength and direction of correlation and Gamma for your readers.** Most of you correctly reported that the correlation between family income and BMI was weak and negative ($r = -0.08$). To be more effective, you should further give an example of what this means, e.g.: “This suggests that as family income increases, youths’ BMI decreases (and vice versa). The relationship is, however, very weak. If we knew youths’
family income, we would only be able to explain less than one percent of the variance in youths’ BMI ($r^2 = 0.0064$).

When interpreting Gamma for the relationship between cigarette smoking and BMI, many of you again reported that the relationship is weak and positive (Gamma = 0.10). To be more effective, go further: “This suggests that as youths’ cigarette smoking increases, so too does their BMI (and vice versa). This relationship is, however quite weak. If we knew youths’ cigarette smoking habits, we would only make about ten percent less errors in predicting BMI than if we knew nothing else about them.”

9. **Is the difference substantively significant?** This is something that is important to keep in front of you. Many of you successfully ran the statistical tests and made the correct conclusion about statistical significance. Don’t forget to go further and interpret whether what you see is *substantively important*. E.g.

**BMI & Region:** It’s a good idea to report what the actual sample means are!

<table>
<thead>
<tr>
<th>REGION</th>
<th>MEAN BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>21.9</td>
</tr>
<tr>
<td>North Central</td>
<td>21.8</td>
</tr>
<tr>
<td>South</td>
<td>22.2</td>
</tr>
<tr>
<td>West</td>
<td>21.7</td>
</tr>
</tbody>
</table>

(note: this mean is calculated from BMI once I deleted BMI values less than 10 and greater than 100 because I thought they were impossible – e.g., a BMI of 9.8 corresponds to a person who is 5’5” and weighs 59 pounds).

If we step back, we see that the above means actually don’t look that different, even though we determined that they were statistically significantly different. For example, take the two that are most extreme – the South and the West – they differ only by (22.2 – 21.7) = 0.5 BMI points. Is this a substantively significant difference? This corresponds to the BMI difference between a person who is 5’5” weighing 133.5 pounds, and a person who is 5’5” weighing 130.5 pounds – not what I would call a huge difference.

Admittedly, this is just taking a single example, and we are comparing averages. Another strategy is to compare the difference to the overall average or overall standard deviation – what is it as a percent of these statistics? Is that a lot or not very much? You want to be able to comment *both* on statistical and substantive significance.

**BMI and Gender.** We can do a similar analysis for these variables.

**BMI and family income; and BMI and cigarette smoothing.** See above under #8 – is the relationship substantively strong?
10. **Important SAS tip:** SAS treats missing values as “infinitely small values.” Thus, the following code would actually result in setting missing values equal to “1” in the new variable:

```sas
data temp;
  set h.pit;
  newbmi=.;
  if ybodm <= 19.13 then newbmi=1;
  if 19.13< ybodm <= 21.03 then newbmi=2;
  if ybodm > 21.03 then newbmi=3;
run;
```

Instead, below is a better way to create this variable: first setting the new variable equal to zero (a value you never really want in this new variable will help you see any coding errors once you print out a frequency table); then the last line cleans up all missing data problems with the new variable by setting it equal to missing if the original variable was missing:

```sas
data temp;
  set h.pit;
  newbmi=0;
  if ybodm <= 19.13 then newbmi=1;
  if 19.13< ybodm <= 21.03 then newbmi=2;
  if ybodm > 21.03 then newbmi=3;
  if ybodm=. then newbmi=.;
run;
```