Homework 8

Make a copy of this document or answer in a separate document.

1. *(1 point) Why is PCA considered a dimensionality reduction technique but not a subset selection technique?*

   Principal Component Analysis is considered a dimensionality reduction technique because it is a popular approach for deriving a low-dimensional set of features from a large set of variables. Instead of choosing a subset of original features, PCA creates a new set of features called principal components.

2. *(1 point) What is a scree plot what does it show? What would you use a scree plot for?*

   A scree plot is a line graph used to visualize and decide how many components or factors to keep in statistical analysis tests such as PCA. It shows the variance a component explains. It helps people decide how many principal components to keep.

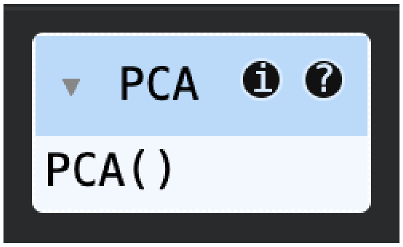
3. *(1 point) What does it mean to prune a Decision Tree? Why would you do this?*

   To prune a decision tree means to remove branches or nodes from the tree that do not provide significant predictive power. This helps reduce overfitting, improve interpretability and reduces complexity.

4. *(1 point) Read [this article]() explain the curse of dimensionality.*

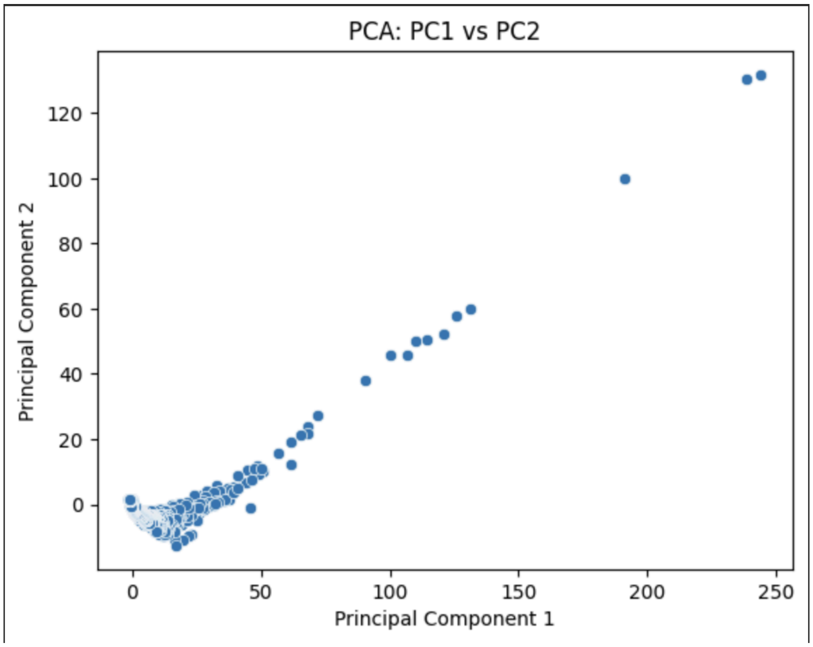
   As the dimensionality of the feature space increases, the number configurations can grow exponentially, and thus the number of configurations covered by an observation decreases. It references increasing data dimensions and its explosive tendencies. Typically, it results in an increase in computational efforts required for its processing and analysis. Some problems that arise is data becoming sparse, distance measures lose meaning, overfitting, underfitting, model performance degrading, Hughes Phenomenon, and computation becoming more expensive.

5. (1 point) `Scale your data and fit your PCA model to the scaled data.`

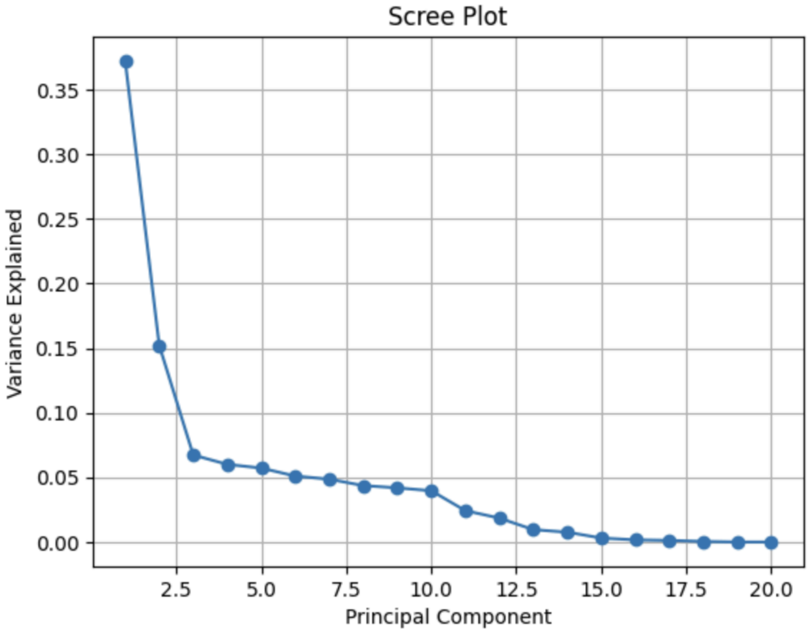


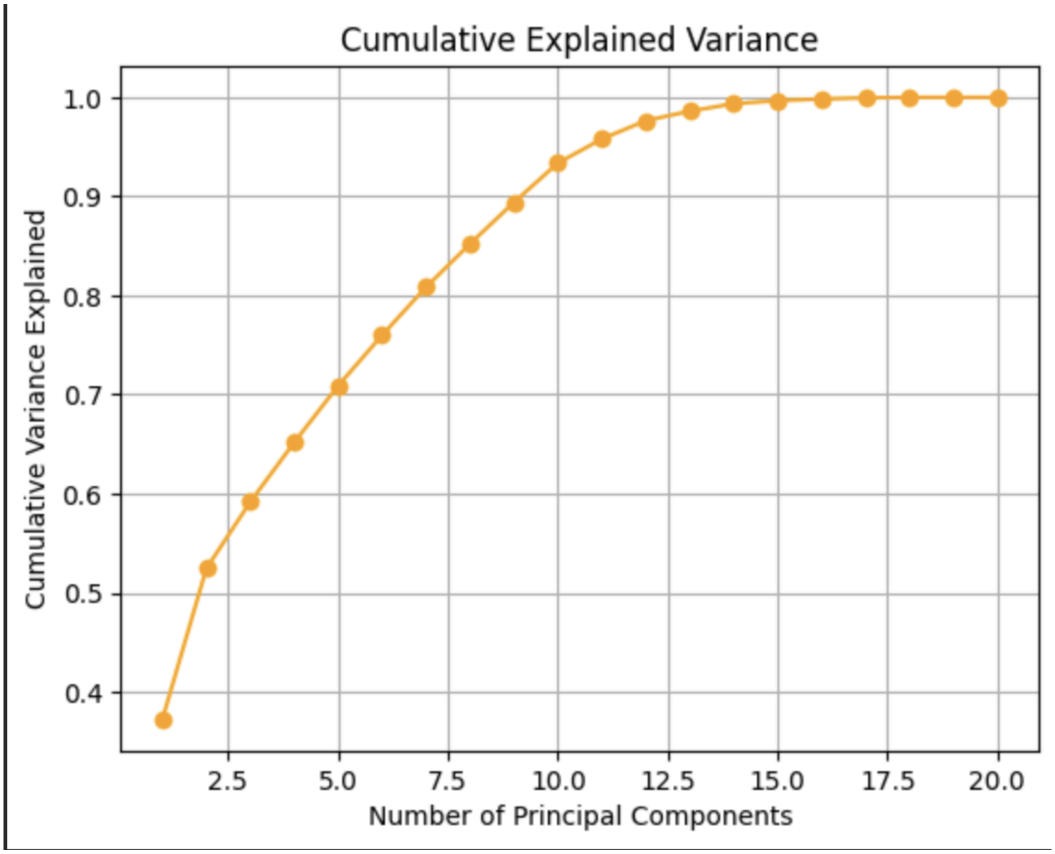
```
PCA()
```

6. (1 point) `Create a scatter plot of PC1 on the x-axis and PC2 on the y-axis`

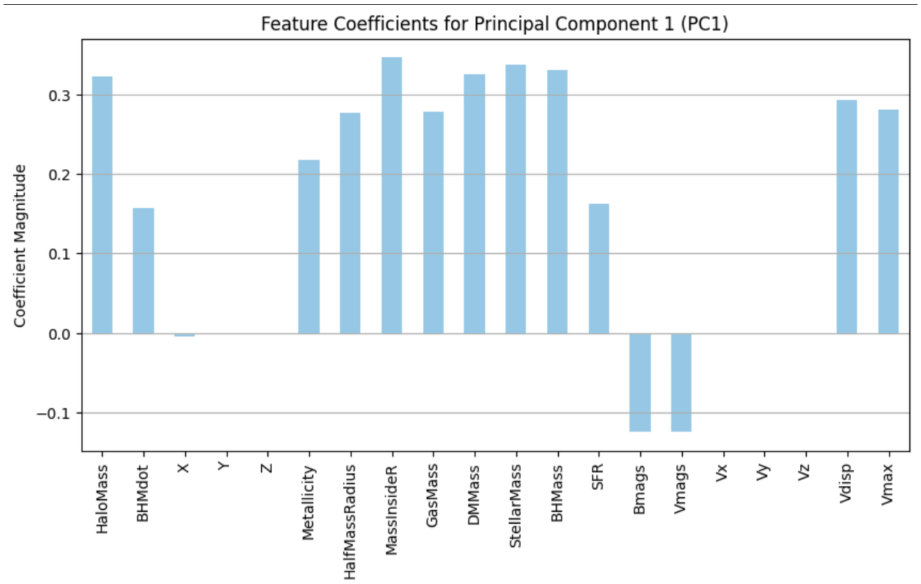


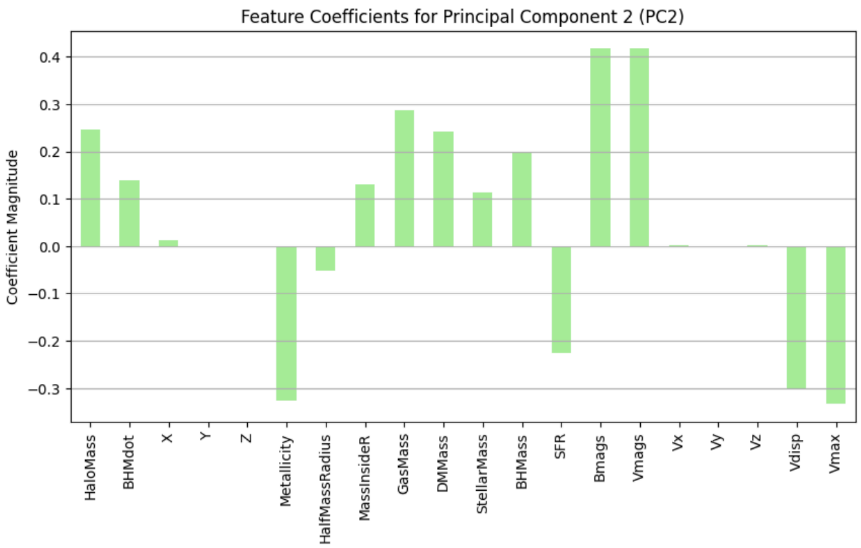
7. (1 point) `Create a scree plot which shows variance explained by each additional principal component. Also create a plot which shows the cumulative explained variance.`

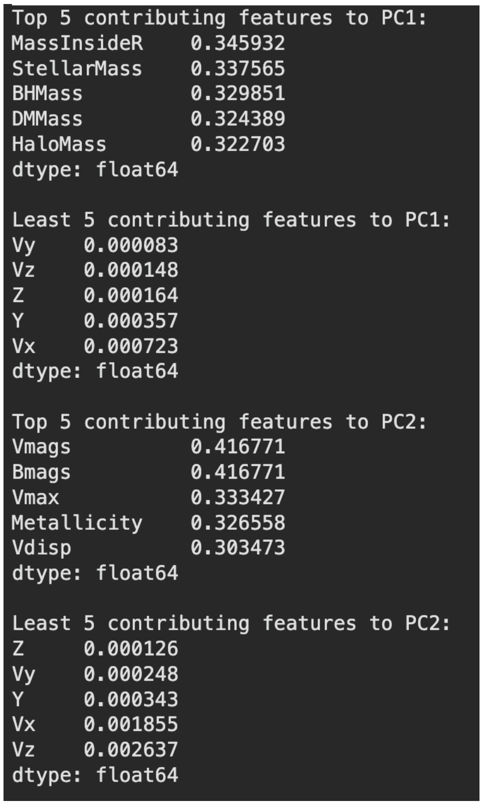




8. (1 point) Create 2 bar charts which show the magnitude of coefficients for PC1 and PC2 respectively the original features. This shows which features were included in each principal component. Which features were important and which were unimportant?







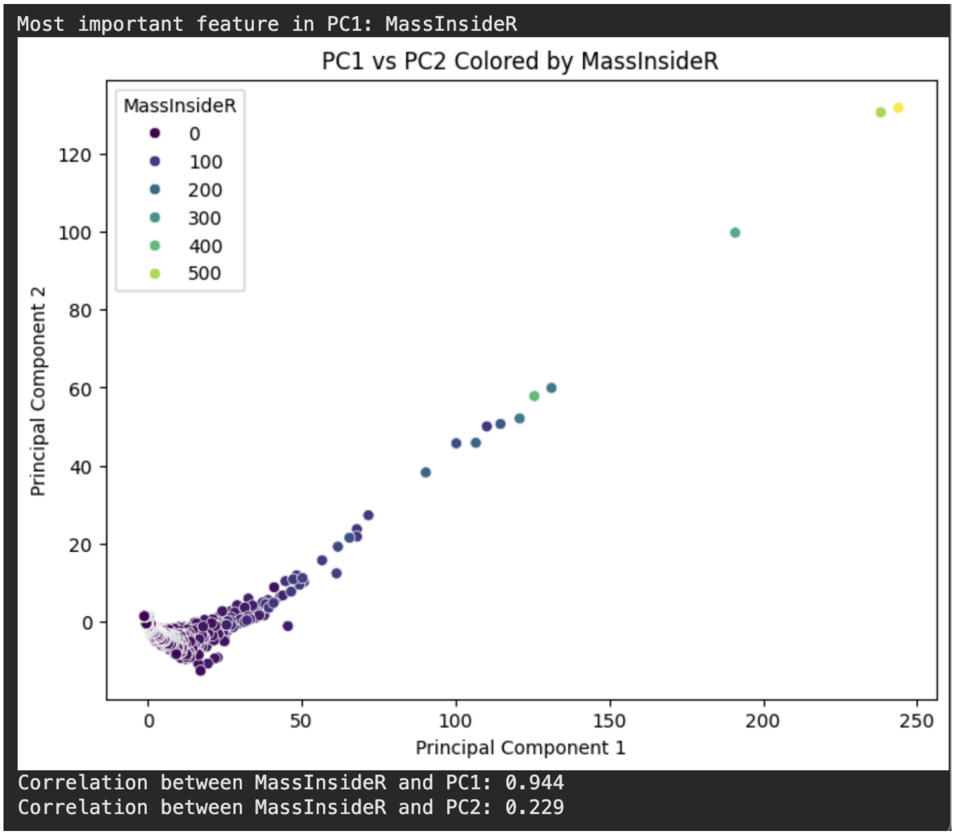
```
Top 5 contributing features to PC1:
MassInsideR    0.345932
StellarMass    0.337565
BHMass         0.329851
DMMass         0.324389
HaloMass       0.322703
dtype: float64

Least 5 contributing features to PC1:
Vy    0.000083
Vz    0.000148
Z     0.000164
Y     0.000357
Vx    0.000723
dtype: float64

Top 5 contributing features to PC2:
Vmags          0.416771
Bmags          0.416771
Vmax           0.333427
Metallicity    0.326558
Vdisp          0.303473
dtype: float64

Least 5 contributing features to PC2:
Z     0.000126
Vy    0.000248
Y     0.000343
Vx    0.001855
Vz    0.002637
dtype: float64
```

9. (1 point) Create a scatter plot of PC1 and PC2 just as before, but color the points based on the most important feature in PC1 (the highest bar from the last question). Does this feature correlate with PC1/PC2 or both?



```
Most important feature in PC1: MassInsideR
```

```
Correlation between MassInsideR and PC1: 0.944
Correlation between MassInsideR and PC2: 0.229
```

The scatter plot shows that the color gradient of MassInsideR changes primarily along the PC1 axis but also slightly along PC2. Numerically, MassInsideR has a strong positive correlation with PC1 and a weaker correlation with PC2. This means MassInsideR is mainly represented by PC1, with a smaller contribution from PC2.

10. (1 point) A much smaller problem to do by hand.
Suppose you do the same task but only with predictors “DMMass” “SFR” and “Metallicity” these tell us how much dark matter there is in the halo, the star formation rate in the halo, and the magnitude of metals in the halo. We will use PCA to construct a new coordinate system. Our first halo has scaled values of :

DMMass = 114.22
SFR = 3.8689
Metallicity = 0.771844

PC1 has a loading vector of [0.4, 0.63, 0.66]
PC2 has a loading vector of [0.91, -0.37, -0.12]
PC3 has a loading vector of [0.11, 0.68, -0.72]

Calculate the *scores* for this datapoint in PC1, PC2, and PC3.

PC1 = (0.4)(114.22)+(0.63)(3.8689)+(0.66)(0.771844) = 48.6348
PC2 = (0.91)(114.22)+(−0.37)(3.8689)+(−0.12)(0.771844) = 102.4161

PC3 = (0.11)(114.22)+(0.68)(3.8689)+(−0.72)(0.771844) = 14.6393

Many problems should be solved using Python. The following notebook has been initialized with the packages and data that you need : link here

To turn in your assignment please take SCREENSHOTS of your results and insert them into (this or another) document for submission. You will then CONVERT THAT DOC TO A .PDF before submitting that. Please embed a hyperlink* to your google notebook INSIDE .pdf that you submit.

* : by “embed the link” I mean that I do NOT want you to copy/paste the disgusting hyperlink. Instead type a word, highlight the word and right click “insert link.” Make sure it works! Then save as .pdf and the link should preserve!

Tran_HW8