Investigation into The Hubble Constant

Herbert S Rowan

Date: Wednesday 8th November

Abstract:

Values for The Hubble constant were obtained from performing linear regression analysis on a dataset. For the whole data set, the value obtained was 82.55 km/s/Mpc. The value of the Hubble constant when filtering the dataset for values of distance <500Mpc was 74.36 km/s/Mpc. The latter value corresponds to literature values with more accuracy, as the mean for the TK method is 75km/s/Mpc. For values >500Mpc, the linear regression model begins to break down.

Introduction:

The Hubble constant, denoted as H0, is a fundamental parameter in cosmology that describes the rate of expansion of the universe. It plays a significant role in understanding the dynamics of our cosmos and the distances to celestial objects. The accurate determination of the Hubble constant is essential for various aspects of astrophysics, from estimating the age of the universe to uncovering the nature of dark energy.

This report investigates the Hubble constant, with a focus on the analysis of two key datasets, 'Ex_Hubble1.csv' and 'Ex_Hubble2.csv.' The objective is to estimate H0 by exploring the relationship between the distance and recessional velocity of objects. The analysis is organized into several components, encompassing data pre-processing, linear regression modelling, and a thorough comparison with published data.

Methodology

The analysis is structured into two main components: linear regression modelling, and comparison with published data.

Linear Regression Modelling:

A linear regression model was employed to investigate the relationship between the distance of galaxies and their corresponding recessional velocity. Linear regression was performed using the scikit-learn library in Python. Two models were developed:

Model 1: Full Dataset Analysis

For the full dataset, we generated a Hubble plot and calculated the Hubble constant using the relationship. The regression equation, R-squared value, and model fit were assessed.

Model 2: Subset Data Analysis

For the subset of data with distances less than 500 Mpc, we performed a secondary linear regression analysis. The same procedure as in Model 1 was applied to estimate H0 for this subset.

Comparison with Published Data

The predicted Hubble constants from both Model 1 and Model 2 were compared with the values reported in 'Ex_Hubble2.csv'. To assess the comparison, the following statistical measures of location and variability were calculated for both sets of predictions and the published data: mean, median, standard deviation and box plots were generated to visualize the distribution of the data.

The comparison included an evaluation of the scientific context and reliability of measurements in the published data.

Results and Discussion

Part A: Simple Regression Models

Initial analysis of the dataset:

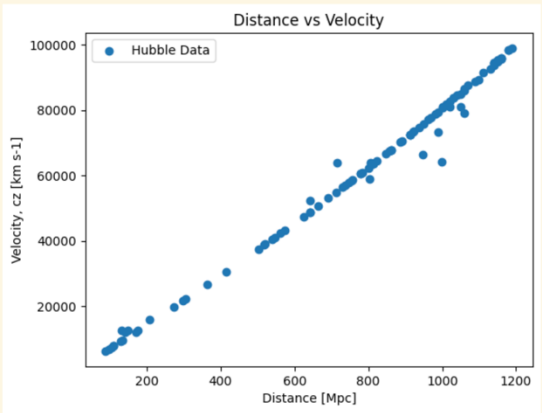


Fig.1: This graph simply shows distance vs velocity for the full data set. From visual inspection it can be seen that there is a slight slope to the data. It appears as though the gradient increases with distance. This is already counter-intuitive to the idea of the Hubble constant as there should be a linear relationship between distance and velocity. Analysis will be continued in its linear properties despite this initial observation.

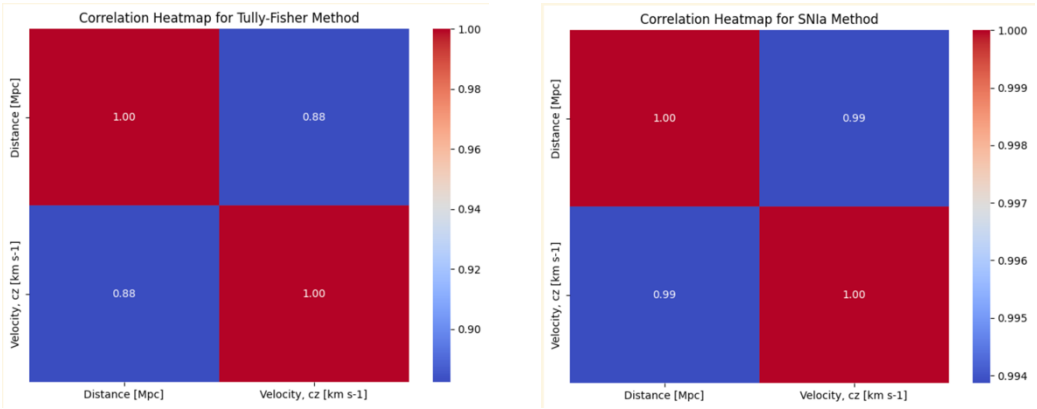


Fig 2. Within 'Ex_Hubble1.csv' there are two methods used to take the measurements: Tully-Fisher and SNIa method. It is clear from the heat maps displayed above that the SNIa method exhibits a higher correlation coefficient of 0.99 as opposed to the Tully-Fisher method that exhibits a correlation coefficient of 0.88. This is a significant difference, so they were investigated further.

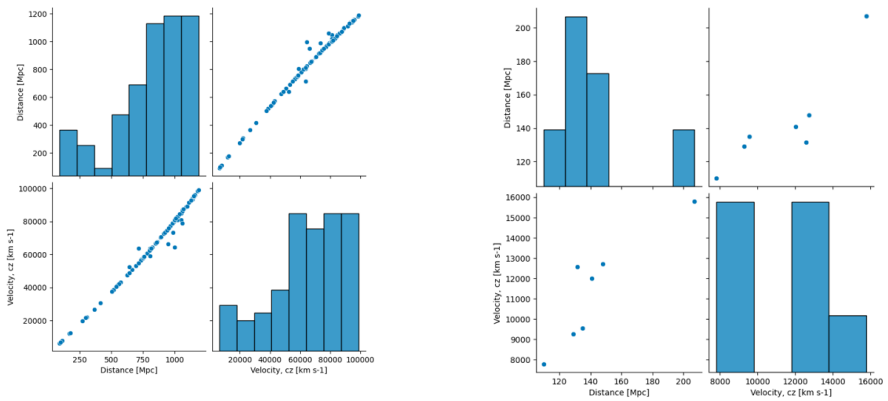


Fig 3. The graph on the left and right show pair plots for the SNIa and Tully-Fisher methods respectively. Upon visual inspection on the SNIa plot the correlation is linear in the subsection 0-500 Mpc, this trend begins to fail at distances of 500+. The Tully-Fisher pair plot upon visual inspection looks linear, however due to the lack of data points it is hard to draw definite conclusions about the efficacy of the method. It would be a valid path to remove this method from the analysis, although due to the ambiguity it will be kept in.

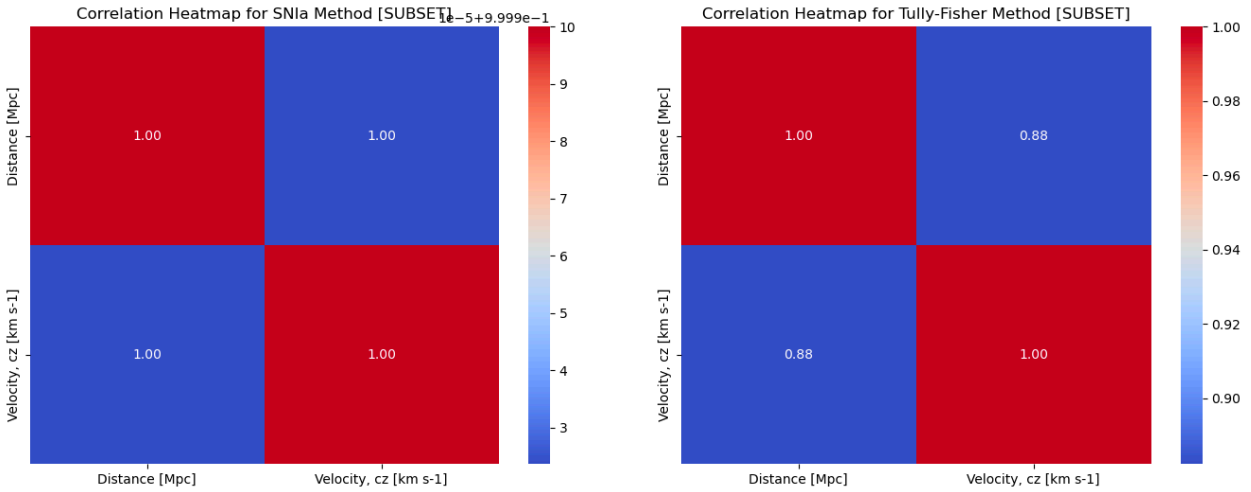


Fig 4. After the observations made in the caption accompanying fig 3, a subset of the data defined by distances < 500Mpc was created in order to verify this observation. The Tully-Fisher method sees very little change as expected due to the lack of data points, however the heatmap for the SNIa sees the correlation coefficient increase from 0.99 to 1.00. Therefore values below 500Mpc exhibit a stronger linear relationship than those above it. It is the data points at larger distances that deviate from the trend. This prompts the necessity for an analysis of two data sets, that of the full data set and a subset of data.

Analysis of the full data set:

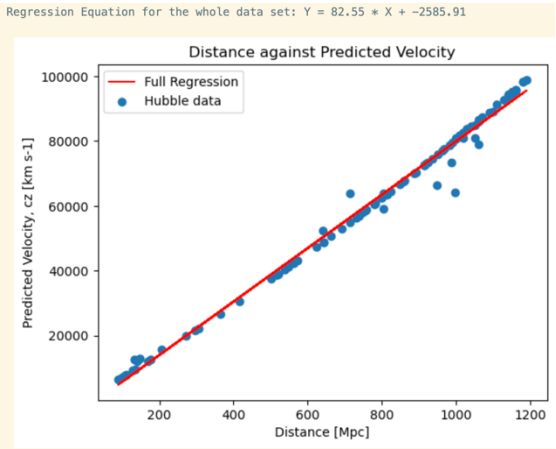


Fig 5. The graph on the left shows the linear regression model for the full data set accompanied by the equation for the respective model. Visually, the fit looks good, however as expected it deviates slightly from the data at values of distance >500Mpc. An R-sqaured value was calculated for this model, being: 0.99011, verifying the assumption that this was a good fit. The value of the Hubble constant obtained from this linear fit is 82.55 km/s/Mpc.

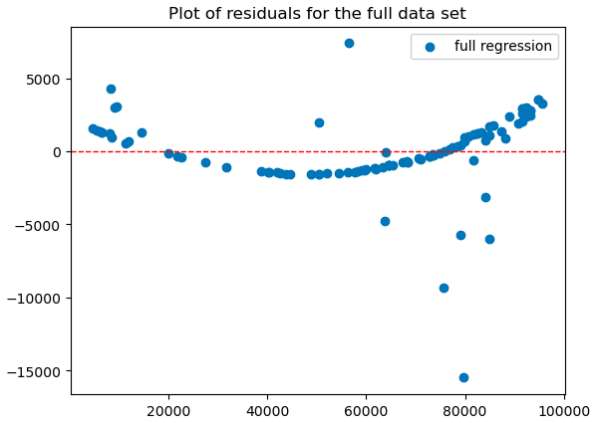


Fig 6. To further investigate this model a plot of the residuals was made. A 'U-shape pattern' or 'funnel shape' is observed. This typically indicates heteroscedasticity, which is a form of non-constant variance of errors or residuals. Heteroscedasticity occurs when the spread or dispersion of the residuals varies systematically across the range of the independent variables. Heteroscedasticity violates one of the assumptions of linear regression, which assumes that the variance of the residuals is constant, this is an assumption called homoscedasticity. The cause of this is the deviation from linearity at distances >500Mpc.

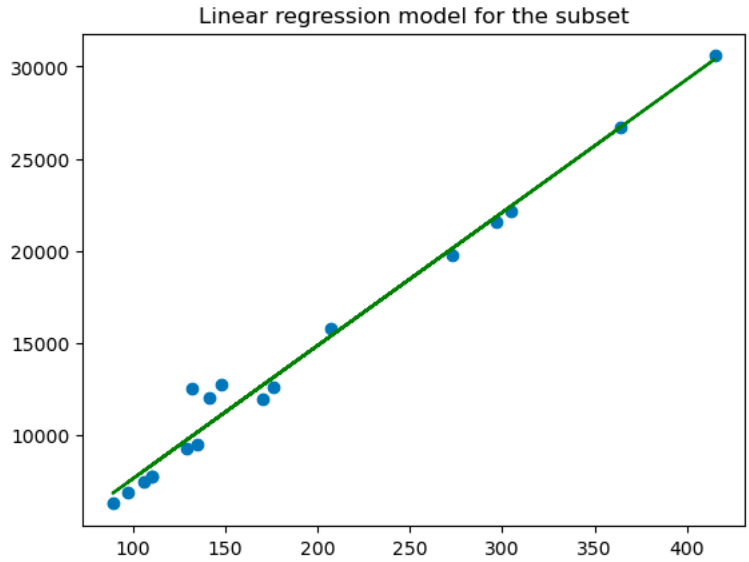


Fig 7. The graph on the left shows the linear regression model for the subset data set (<500Mpc). As expected, the regression model does not deviate significantly from the data. The group of three points that deviate from the regression at 150 are essentially outliers from the model so this was inspected further. The value obtained for the Hubble constant from this regression model is: 72.19 km/s/Mpc.

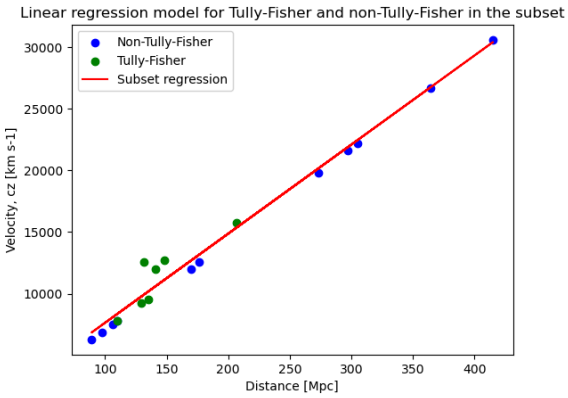
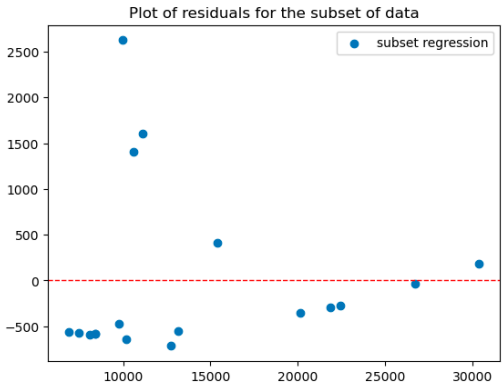


Fig 8. The graph on the left shows the linear regression model for the subset data set. As expected, the three points discussed in fig 7 belong to the Tully-Fisher method, otherwise the points lie very close to the regression model. The R-squared value when of the whole subset is 0.98293. When the Tully-Fisher data is removed the R-squared score is 0.99985, which is better than the score of 0.99011 for both methods on the full data set as shown in fig 5.



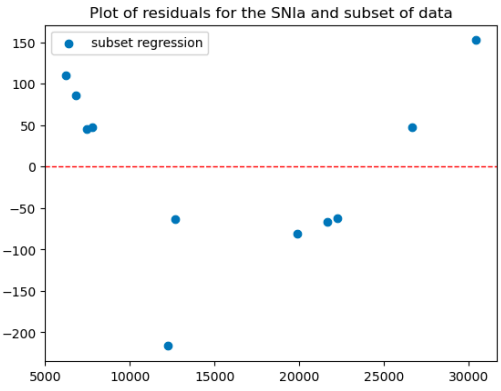


Fig 9. The graph on the left shows the residuals plot for the subset of the data with all methods included, you can see the range of values is approximately 3000. The graph on the right shows the plot of residuals for the subset of the data that only includes the SNIa data. The range of values in this graph is approximately 350, this is approximately an order of magnitude smaller than the range from the graph on the right. Therefore, as expected we see a much more accurate fit of the linear regression model when the Tully-Fisher data is removed.

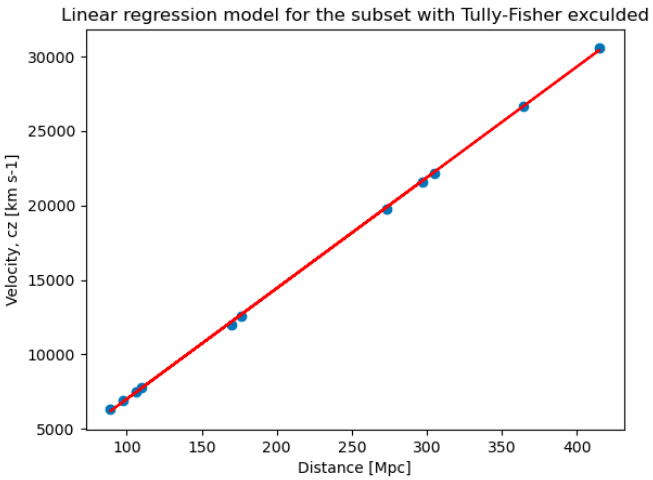


Fig 10. The graph on the right shows the linear regression model for the subset of data with Tully-Fisher data removed. The value of the Hubble constant obtained from this regression model is: 74.36 km/s/Mpc. Thus far in the report we have obtained three values of the Hubble constant from analysing the data from different approaches, more discussion on the ideal value will follow later in the report.

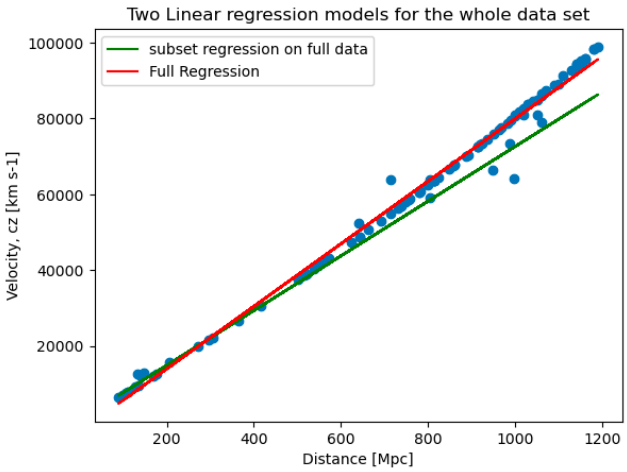


Fig 11. The graph on the right shows two of the linear regression models, one for the subset of the data, continued on through the rest of the data and the other the regression model for the full data set plotted together. It can be seen how the two models deviate from one another.

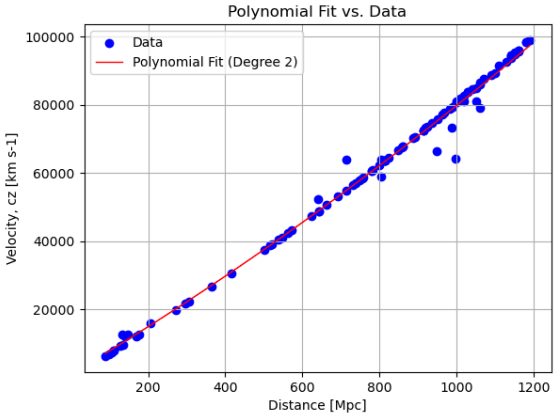
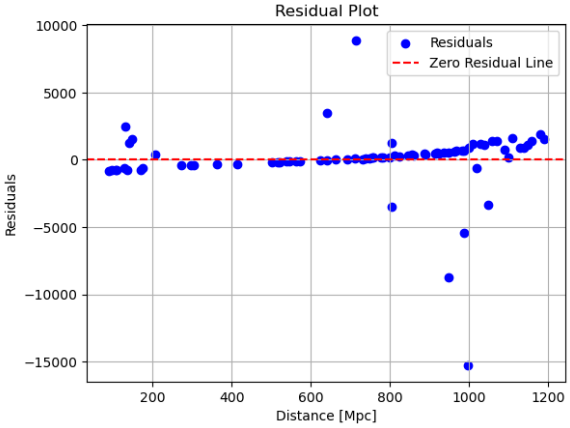


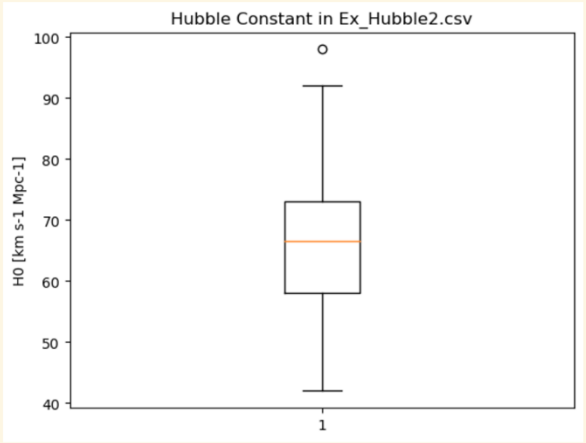
Fig 12. The graph on the right shows a polynomial fit of the 2nd degree on the full data set using the numpy module.

Fig 13. The graph on the right shows the residual plot of the polynomial fit. It can be seen that it is a good fit apart from a few outliers, this could point to the potentially non-constant nature of the Hubble constant, however this hypothesis would merit further investigation. Although even with the polynomial fit, errors increase for values >600Mpc, and increase in magnitude as distance increases.



Part B: Statistical Measures of Location and Variability

This sub-section of the results and discussion section focuses on the analysis of 'Ex_Hubble2.csv'. This csv file contains data related to the Hubble constant measured by various sources. Each row represents a different measurement, and each column provides information on the measured value of H0, positive and negative errors, the date of measurement, the type of measurement and the source of the measurement which contains bibliographic information. Using this data source, we will be able to compare the results of the data in 'Ex_Hubble1.csv' to the data of the broader scientific community.




```
Statistical Measures for Hubble Constant in Ex_Hubble2.csv:
Mean: 66.29 km/s/Mpc
Standard Deviation: 11.18 km/s/Mpc
Median: 66.50 km/s/Mpc
```

Fig 14. The graph on the left shows a box plot made from all the values of the Hubble constant in the dataset. An outlier is shown by the circle. The range, median, mean, standard deviation and IQR can be seen from the box plot and corresponding data below. The three values of the Hubble constant obtained from part A are 74.36, 72.19 and 82.55 km/s/Mpc. The value of 82.55 exists more than one standard deviation away from the mean whereas the other two values fall within it. 72.19 is the value that is closest to the mean.

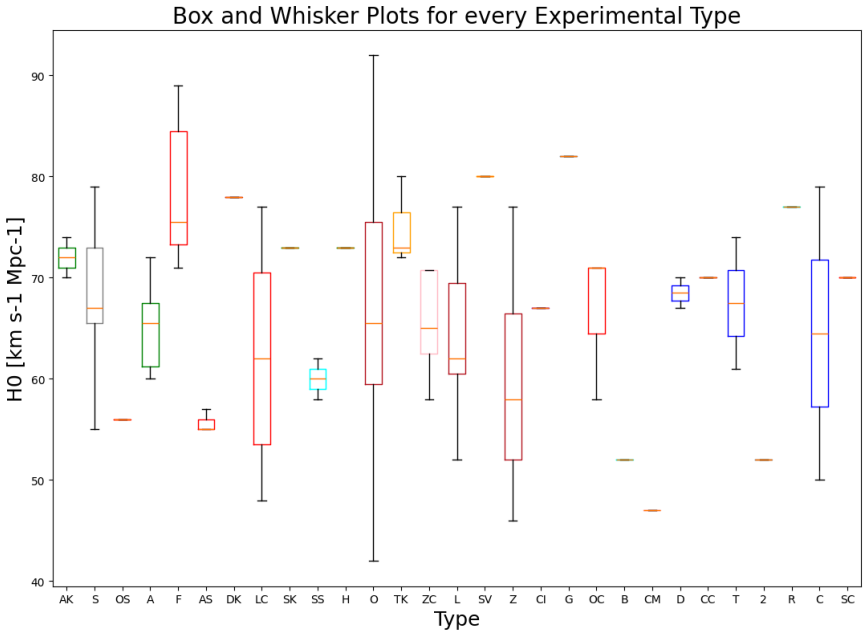
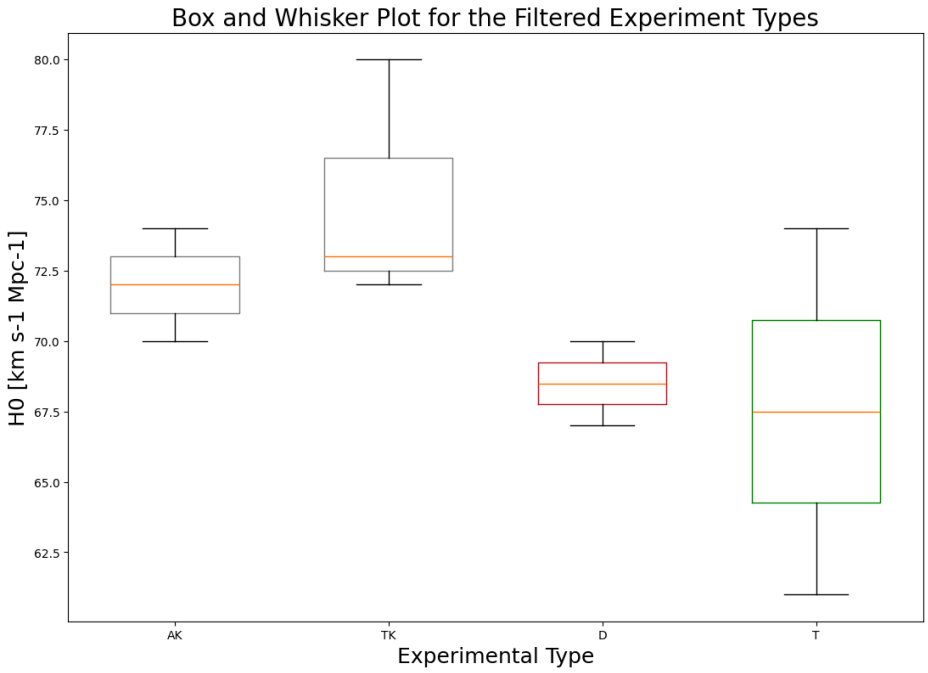


Fig 15. The graph on the left shows multiple box plots made for each of the recording methods. This graph was produced to investigate the potentially unreliable recording methods used in the data set. It can be seen that certain methods such as O have vast ranges while others have significant outliers such as L. This prompted the development of an algorithm to filter out the data which is skewing the results.




```
Mean for AK: 72.00
Mean for TK: 75.00
Mean for D: 68.50
Mean for T: 67.50
Standard Deviation for AK: 2.00
Standard Deviation for TK: 4.36
Standard Deviation for D: 2.12
Standard Deviation for T: 9.19
The average mean for the filtered data set is 70.75
The average standard deviation for the filtered data set is 4
```

Fig 16. The graph on the left shows multiple box plots made for each of the recording methods that fit within the range 59 to 81 because these values fall into the space defined by ~1 standard deviation away from the mean as defined in fig 14, and the corresponding data. The average mean for the new data set agrees with the data from part A more than the unfiltered data set box plots. Recording methods such as TK are very close to 74.36 and AK is very close to 72.19. In the light of this new data set, the value of 82.55 still does not perform well for the same reasons stated in the caption for fig 14.

Conclusion and further discussion

It is acknowledged that limitations in the analysis, including the assumption of a linear relationship and potential uncertainties in measurements make it a challenge to determine the exact value of the Hubble constant. Further research is needed to address these limitations and refine our understanding of the Hubble constant. This analysis contributes to the ongoing exploration of the Hubble constant and the expanding universe.

However, the estimates provided by our linear regression models align closely with published data, indicating the potential for reliable predictions. Especially the values obtained for the Hubble constant when considering distances below 500Mpc. This is indicated by the recording methods such as TK being very close to 74.36 and AK being very close to 72.19. Thus, in light of scientific context, the confidence that one can have in the measurements taken in part A increases. The high correlation coefficients such as 1.00 for distances below 500Mpc and with the Tully-Fisher method excluded, and R-squared values obtained in part A such as 0.99985 (again for distances below 500Mpc and with the Tully-Fisher method excluded) further strengthen this conclusion. This conclusion becomes

significantly more unreliable when considering distances above 500Mpc. The R-squared value drops from 0.99985 to 0.99011 and the correlation coefficient drops form 1.00 to 0.99 when you include Tully-Fisher and values greater than 500Mpc. The residuals also dramatically increase. The Hubble constant becomes more difficult to measure at large distances, some examples being the influence of dark energy, cosmic variance and general relativity however the nature of these phenomena are beyond the scope of this report.