**1. Introduction**

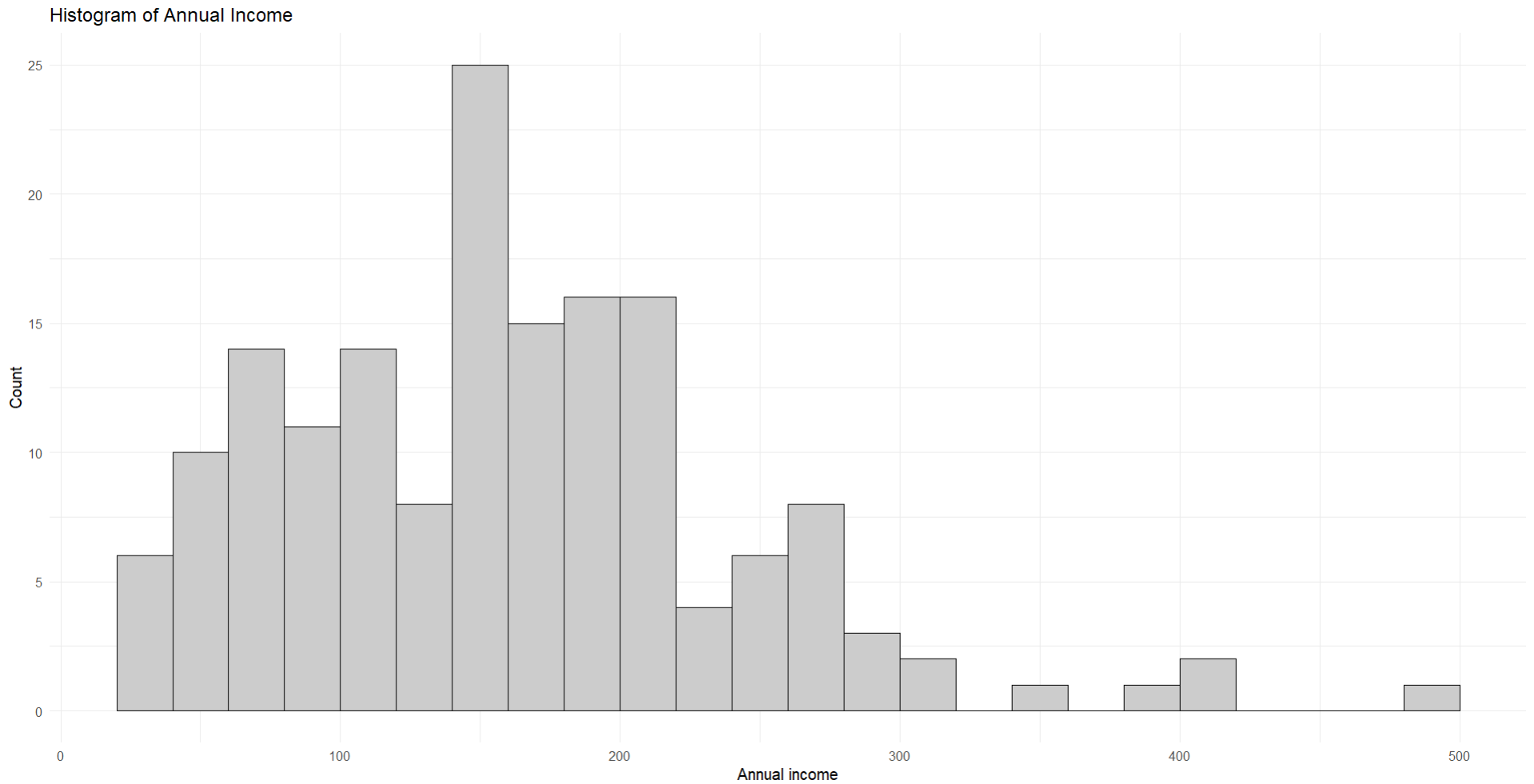
* Assignment Topic: Which factors are associated with annual income in this Dataset?
* Variables studied: Age, Education, Experience, Family Size, Farm Size, Access to Info Sources, Knowledge, Problem (and Annual Income as dependent variable).

**2. Data & Methods**

* Methods used:
  + Descriptive statistics for annual income (mean, SD, min, max).
  + Pearson correlation (r) between annual income and each numeric predictor; report r and p-value.
  + Income and age were categorized into **three groups (tertiles)**: Low / Medium / High and Young / Middle / Old respectively.
  + Group summaries (mean and SD) by income tertile and by age tertile were computed.
  + Visuals: histogram, correlation heatmap, boxplots, and scatter with fitted line.

**3. Descriptive results — Annual Income (k)**

* **Mean (Annual Income)** = **162.362** and **SD** = **82.31**
* **Min** = 25 and **Max** = 500

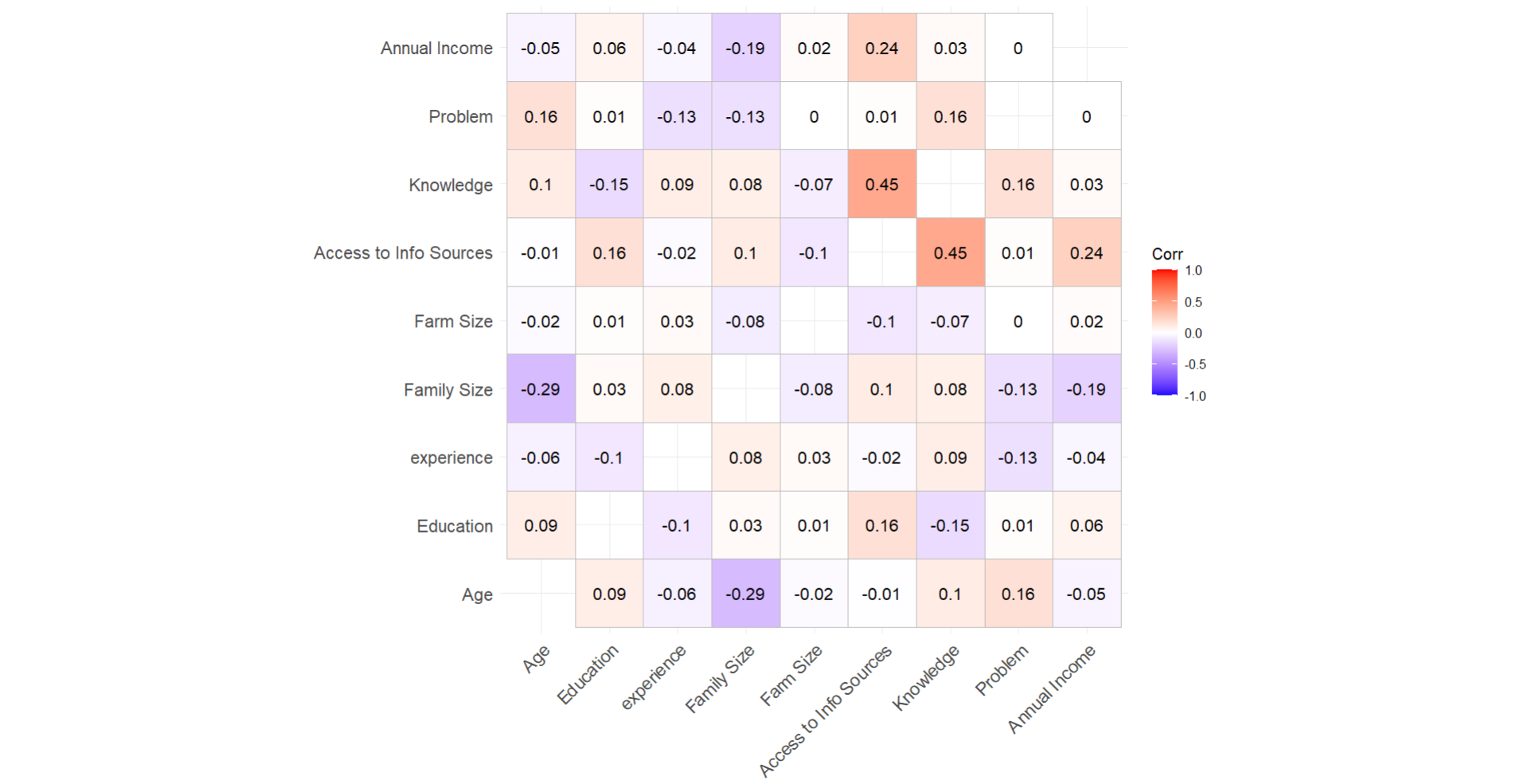


**Figure 1:** *Distribution of annual income.*

**4. Correlations with Annual Income (Pearson r and p-value)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictor** | **Pearson r** | **p-value** | **Interpretation** |
| Age | -0.0479 | 0.5436 | no meaningful correlation |
| Education | 0.0617 | 0.6116 | weak, not significant |
| Experience | -0.0421 | 0.5934 | weak, not significant |
| **Family Size** | **-0.1925** | **0.0138** | small negative, **significant** |
| Farm Size | 0.0155 | 0.8446 | no correlation |
| **Access to Info Sources** | **0.2373** | **0.0023** | small–moderate positive, **significant** |
| Knowledge | 0.0305 | 0.6996 | no correlation |
| Problem | 0.0009 | 0.9905 | no correlation |

* **Main takeaways from correlations:** only *Access to Info Sources* (positive) and *Family Size* (negative) show statistically significant associations with income at α = 0.05. Other variables show weak or no association.



**Figure 2 (Correlation heatmap):** *Correlation matrix (Pearson r) for numeric variables.*

**5. Income categories**

* Counts in each income tertile:
  + **Low** = 55 people
  + **Medium** = 54 people
  + **High** = 54 people

**Group summary (means and SDs)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Income group** | **Mean Annual Income** | **SD (Income)** | **Mean Age** | **Mean Education** | **Mean Experience** | **Mean Access** | **Mean Knowledge** | **Mean Family Size** |
| Low | 78.327 | 26.621 | 42.473 | 3.355 | 24.345 | 13.545 | 12.436 | 6.636 |
| Medium | 158.741 | 17.678 | 42.870 | 3.586 | 23.667 | 13.463 | 12.111 | 6.667 |
| High | 251.574 | 64.985 | 40.852 | 3.600 | 23.537 | 15.778 | 12.907 | 6.185 |

* **Interpretation:**
  + The High-income group has noticeably higher mean **Access to Info Sources** (15.78) than Low/Medium groups (~13.5).
  + Family size slightly smaller in High group (6.19) compared with Low (6.64).
  + Education differences are small across groups in this dataset.

**6. Age groups (tertiles) vs Income — summary**

Counts:

* Young = 62, Middle = 53, Old = 48

Mean income by age group:

|  |  |  |
| --- | --- | --- |
| **Age group** | **Mean Income** | **SD Income** |
| Young | 163.984 | 80.061 |
| Middle | 172.811 | 85.902 |
| Old | 148.729 | 80.945 |

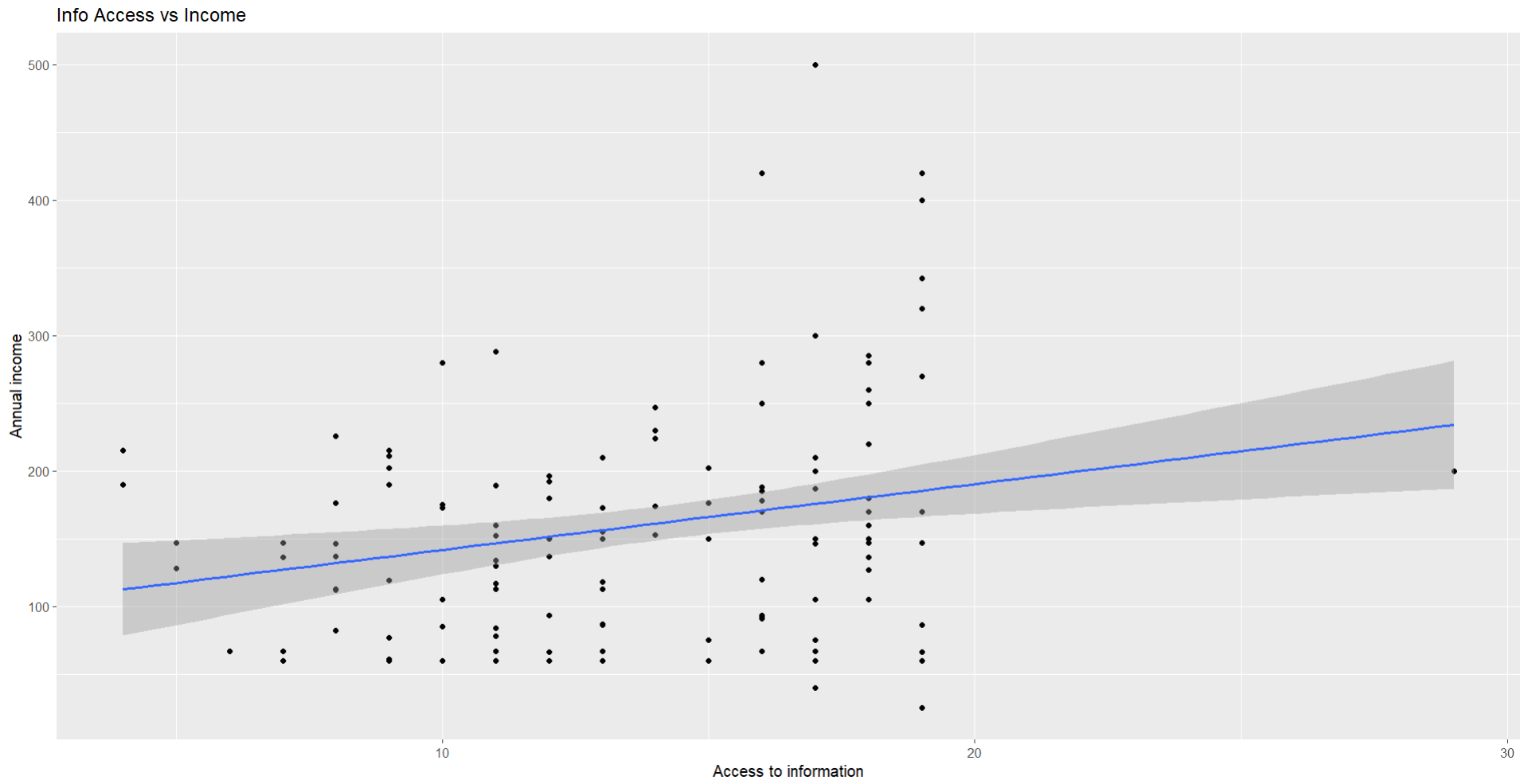
* **Interpretation:** middle-aged group shows slightly higher mean income than young and old, but differences are modest.

**7. Results**

* **Descriptive:**  
  “In this sample (N = 163), mean annual income = **162.36** (SD = **82.31**), ranging from 25 to 500.”
* **Correlation summary:**  
  “Pearson correlations show that **Access to Info Sources** has a positive association with income (r = **0.237**, p = **0.0023**). **Family Size** has a negative association with income (r = **−0.193**, p = **0.0138**). Other variables (Age, Education, Experience, Farm Size, Knowledge, Problem) show weak or non-significant correlations with income.”
* **Income tertiles:**  
  “Income tertiles divide the sample into Low (n=55), Medium (n=54), and High (n=54). Mean income in these groups are 78.33 (Low), 158.74 (Medium), and 251.57 (High). The High-income group shows higher average access to information and slightly smaller family size.”
* **Age groups:**  
  “Mean income by age tertile is 164 (Young), 173 (Middle), and 149 (Old). Differences are small.”

**8. Interpretation and short discussion**

* People with **more access to information** tend to have higher income. This could mean they get better market info, job leads, or farming advice that increases earnings.
* **Larger family size** is associated with lower income. This may reflect resource dilution (same total income spread over more family members) or social-economic differences.
* **Education** and **knowledge** in this dataset do not show strong direct correlations with income; possible reasons: limited variation in measures, measurement error, or confounding (education relates to access to info).
* **Caveat:** These are **associations** from cross-sectional data. We cannot claim cause → effect.

 **Figure 3 (Scatter + lm):** *Access info vs Annual Income with linear fit.*

**9. Conclusion & recommendations**

* Improving **access to information** could be a feasible intervention to help increase incomes.
* Further research: multiple regression to control confounders, and longitudinal data to test causal effects.

**10. Appendix**

The code used to analysis this report-

#First export the data form the local drive to R

#descriptive of income

student\_data %>% summarise(

mean\_income = mean(`Annual Income`,na.rm= TRUE),

sd\_income = sd(`Annual Income`,na.rm= TRUE),

max = max(`Annual Income`,na.rm= TRUE),

min = min(`Annual Income`,na.rm= TRUE)

)

#calculating the Pearson correlation and p value

library(Hmisc)

res <- rcorr(as.matrix(student\_data), type="pearson")

r\_mat <- res$r

p\_mat <- res$P

#make this into tidy format for the analysis

r <- r\_mat["Annual Income", ]

p <- p\_mat["Annual Income", ]

corr\_income <- tibble(

variable = colnames(r\_mat),

coorelation = r,

p\_value = p) %>% filter(variable != "Annual Income") #here it is used to make the the age variable disappear

#Coorelation heatmap

library(ggcorrplot)

corr\_all <- cor(student\_data, use="pairwise.complete.obs")

ggcorrplot(corr\_all, lab = TRUE, show.diag = FALSE)

#adding more plot for the explination

ggplot(student\_data, aes(`Access to Info Sources`,`Annual Income`)) +

geom\_point() +

geom\_smooth(method="lm", se=TRUE) +

labs(x="Access to information", y="Annual income", title="Info Access vs Income")

#age categories adding in the student data

student\_data <- student\_data %>%

mutate(age\_cat = ntile(`Age`, 3)) %>%

mutate(age\_cat= recode(age\_cat,

`1` = "Young",

`2` = "Middle",

`3` = "Old"

))

#income categories adding in the student data

student\_data <- student\_data %>%

mutate(income\_cat = ntile(`Annual Income`, 3)) %>%

mutate(income\_cat= recode(income\_cat,

`1` = "Low",

`2` = "Medium",

`3` = "High"

))

student\_data %>% count(age\_cat)

student\_data %>% count(income\_cat)

#table for just analysis

New\_table <- student\_data %>% group\_by(income\_cat) %>%

summarise(

mean\_age = mean(`Age`, na.rm=TRUE),

mean\_education = mean(`Education`, na.rm=TRUE),

mean\_experience = mean(experience, na.rm=TRUE),

mean\_access = mean(`Access to Info Sources`, na.rm=TRUE),

mean\_knowledge = mean(`Knowledge`, na.rm=TRUE),

mean\_family = mean(`Family Size`, na.rm=TRUE)

)

#histogram plot

ggplot(student\_data, aes(`Annual Income`)) +

geom\_histogram(binwidth = 20, boundary = 0, color = "black", fill = "grey80") +

labs(title = "Histogram of Annual Income",

x = "Annual income",

y = "Count") +

theme\_minimal()