exploratory_modelling

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("ggplot2")
library("patchwork")
```

```
ds_exp <- read.csv("~/lab-2-team-no-1-s/data/interim/GSS_exploration_set.csv")
```

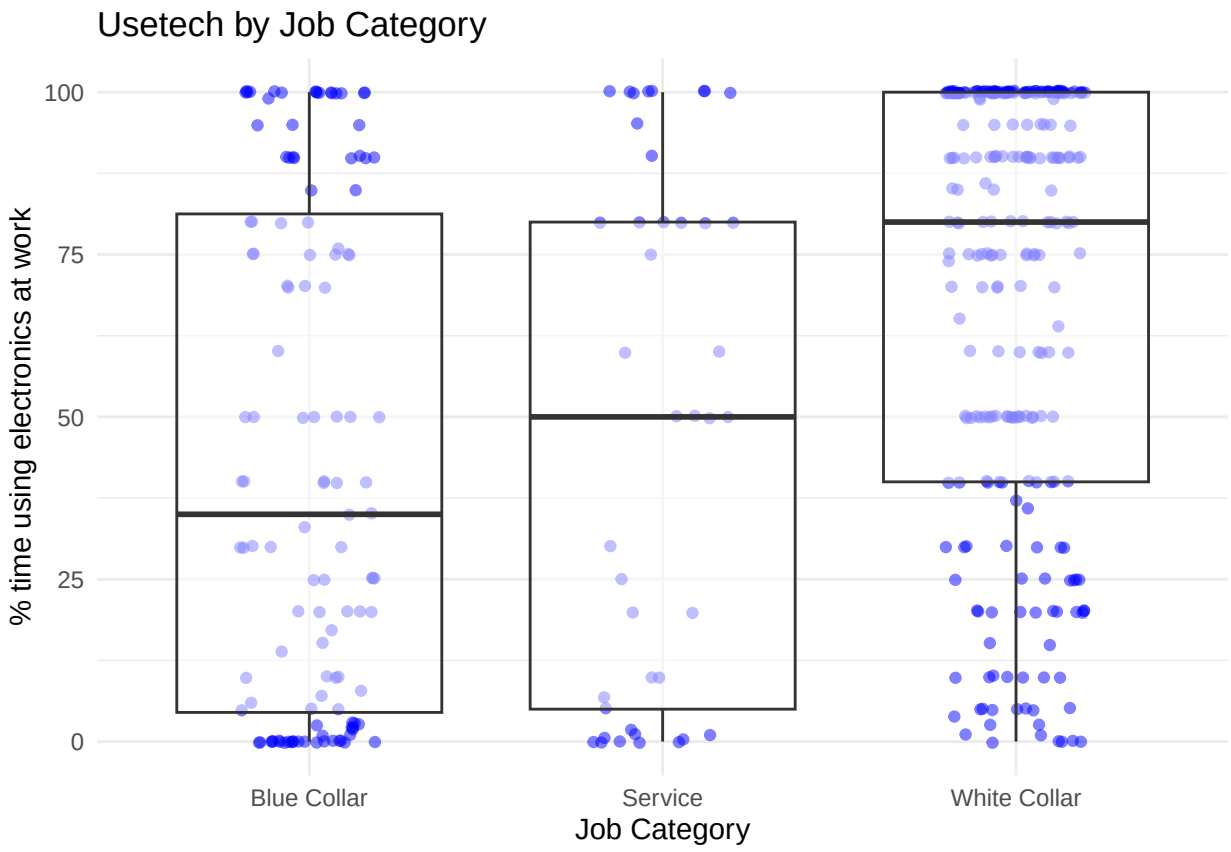
```
ds_exp$industry <- as.factor(ds_exp$industry)
ds_exp$job_ctg <- as.factor(ds_exp$job_ctg)
```

```
## Usetech Distribution

usetech_by_job_plot <- ds_exp %>%
                        ggplot()+
                        aes(x=job_ctg, y=usetech)+
                        geom_jitter(width=0.2, height=0.2, color="blue", alpha=0.5)+
                        geom_boxplot(outlier.shape = NA, alpha=0.5)+
                        labs(x="Job Category", y="% time using electronics at work", title = "Usetech by
                        theme_minimal()


usetech_facet_by_age <- ds_exp %>%
                          mutate(age_group = cut(age, breaks = c(10, 20, 30, 40, 50, 60, 70, 80))) %>%
                          ggplot(aes(x = job_ctg, y = usetech)) +
                          geom_jitter(width = 0.2, height = 0.2, alpha = 0.5, color = "blue") +
                          geom_boxplot(outlier.shape = NA, alpha = 0.3) +
                          facet_wrap(~ age_group) +
                          labs(x = "Job Category", y = "% Time Using Electronics at Work",
                               title = "Use of Technology by Job Category and Age Group") +
                          theme_minimal()

usetech_histogram <- ds_exp %>%
                        ggplot() +
                        aes(x = usetech) +
                        geom_histogram(binwidth = 10, fill="skyblue", color="black") +
                        geom_density(aes(y = after_stat(count) * 10), color = "darkred", size = 1) +
                        scale_x_continuous(breaks = seq(0, 100, by = 10)) +
                        labs(x = "% Time Using Electronics at Work", y = "Count", title = "Distribution
```

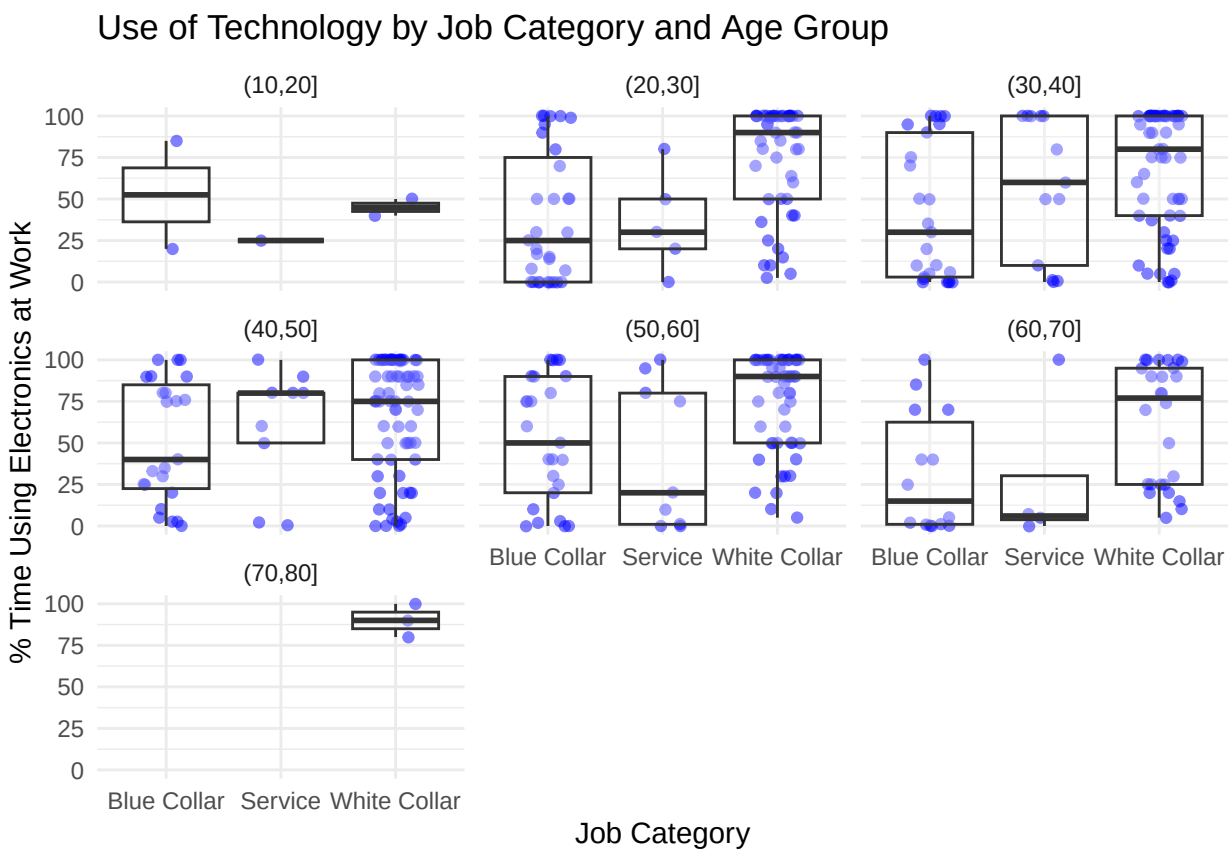
```
                        theme_classic()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

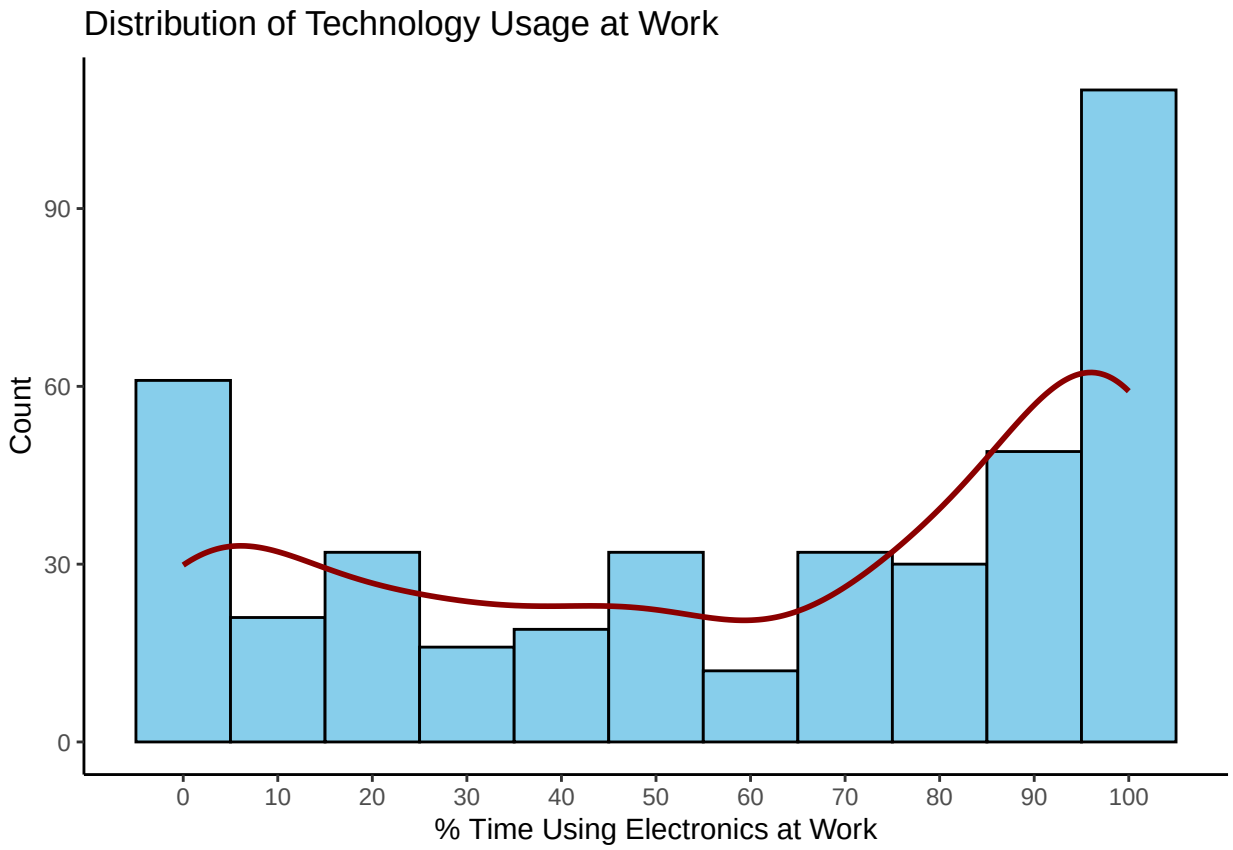
```
usetech_by_job_plot
```




```
usetech_facet_by_age
```



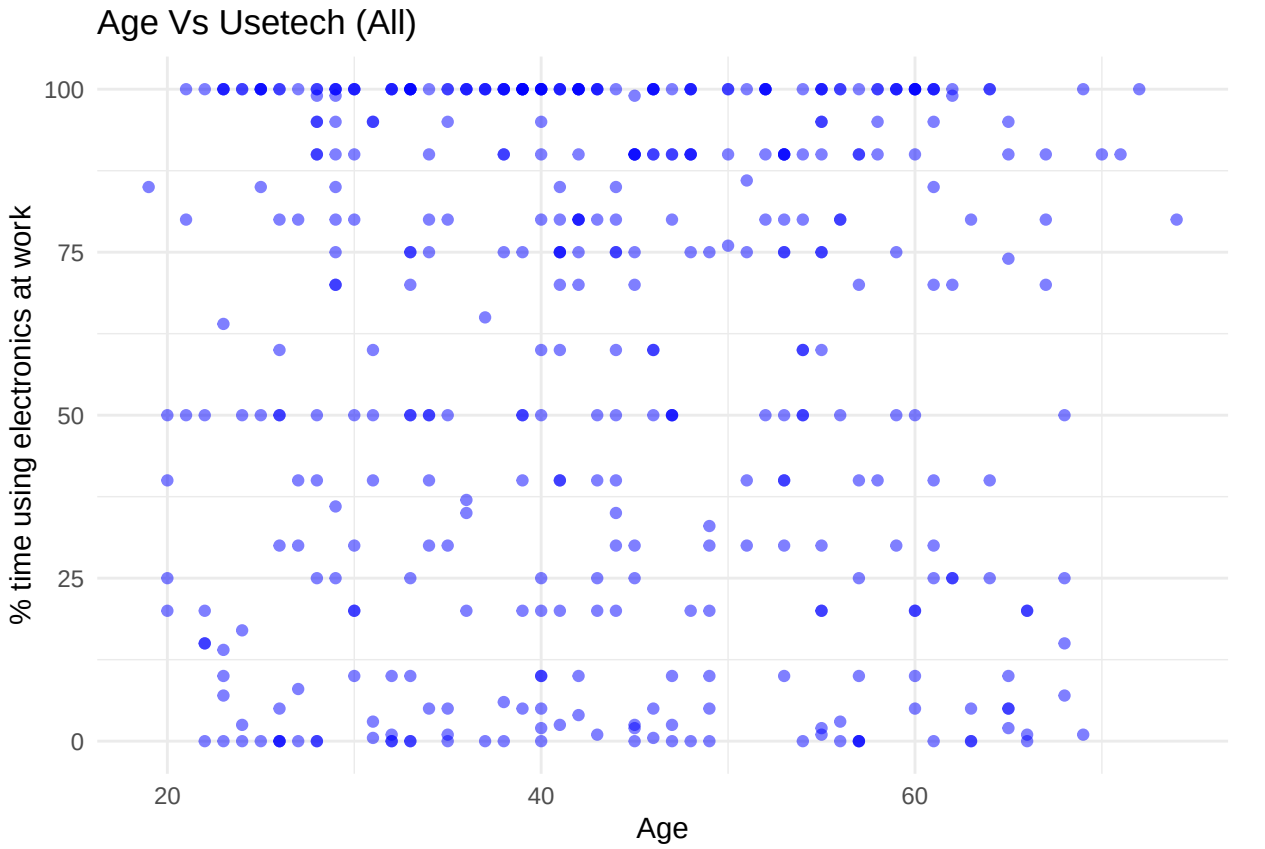

```
usetech_histogram
```



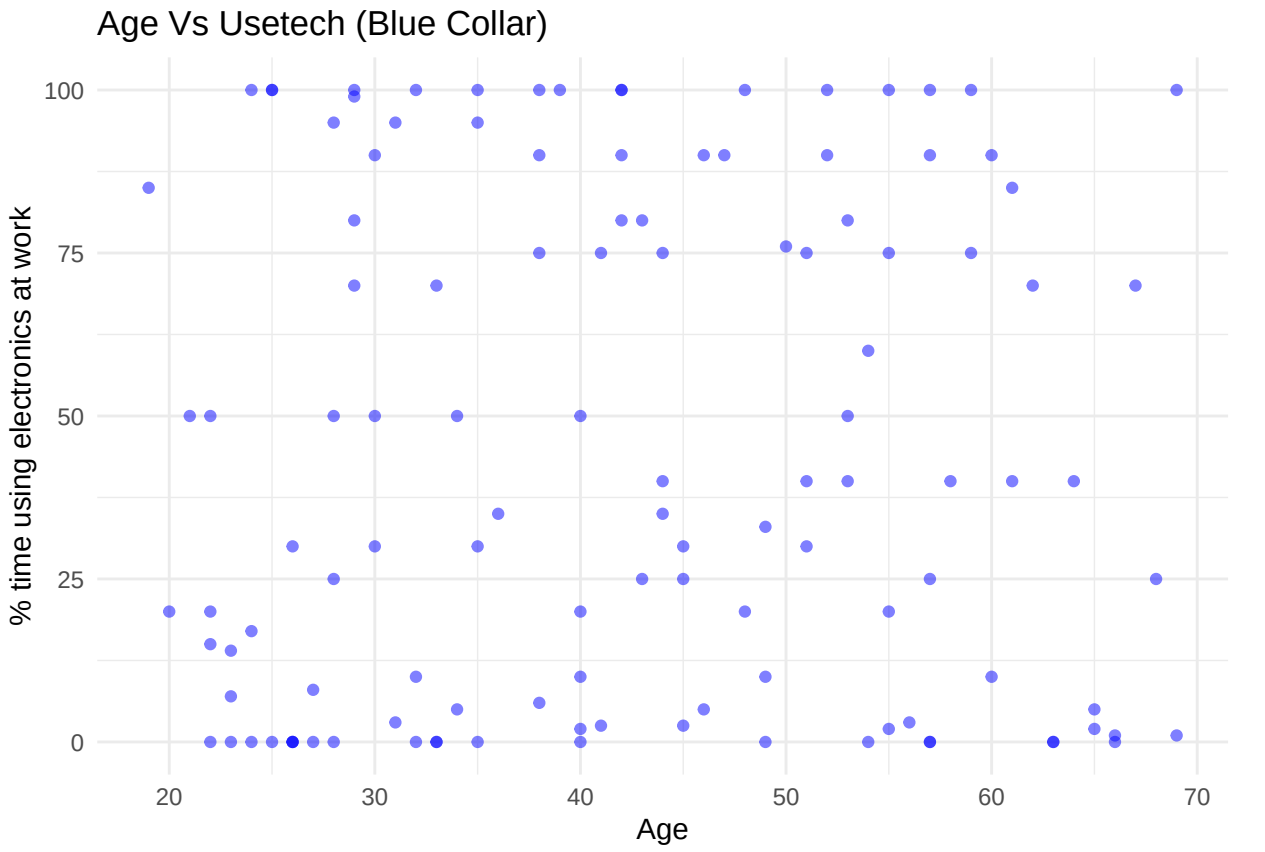

```
## Save the plots.
ggsave("~/lab-2-team-no-l-s/notebooks/plots/usetech_by_job_plot.png", plot = usetech_by_job_plot, width
ggsave("~/lab-2-team-no-l-s/notebooks/plots/usetech_facet_by_age.png", plot = usetech_facet_by_age, wid
ggsave("~/lab-2-team-no-l-s/notebooks/plots/usetech_histogram.png", plot = usetech_histogram, width = 8

## Age Distribution
```

```
age_vs_usetech_scatter <- ds_exp %>%
                          ggplot()+
                          aes(x=age, y=usetech)+
                          geom_point(color="blue", alpha=0.5)+
                          labs(x="Age", y="% time using electronics at work", title = "Age Vs Usetech
                          theme_minimal()

age_vs_usetech_bluecollar_scatter <- ds_exp %>%
                                     filter(job_ctg=="Blue Collar") %>%
                                     ggplot()+
                                     aes(x=age, y=usetech)+
                                     geom_point(color="blue", alpha=0.5)+
                                     labs(x="Age", y="% time using electronics at work", title = "Ag
                                     theme_minimal()

age_vs_usetech_whitecollar_scatter <- ds_exp %>%
                                     filter(job_ctg=="White Collar") %>%
                                     ggplot()+
                                     aes(x=age, y=usetech)+
                                     geom_point(color="blue", alpha=0.5)+
```

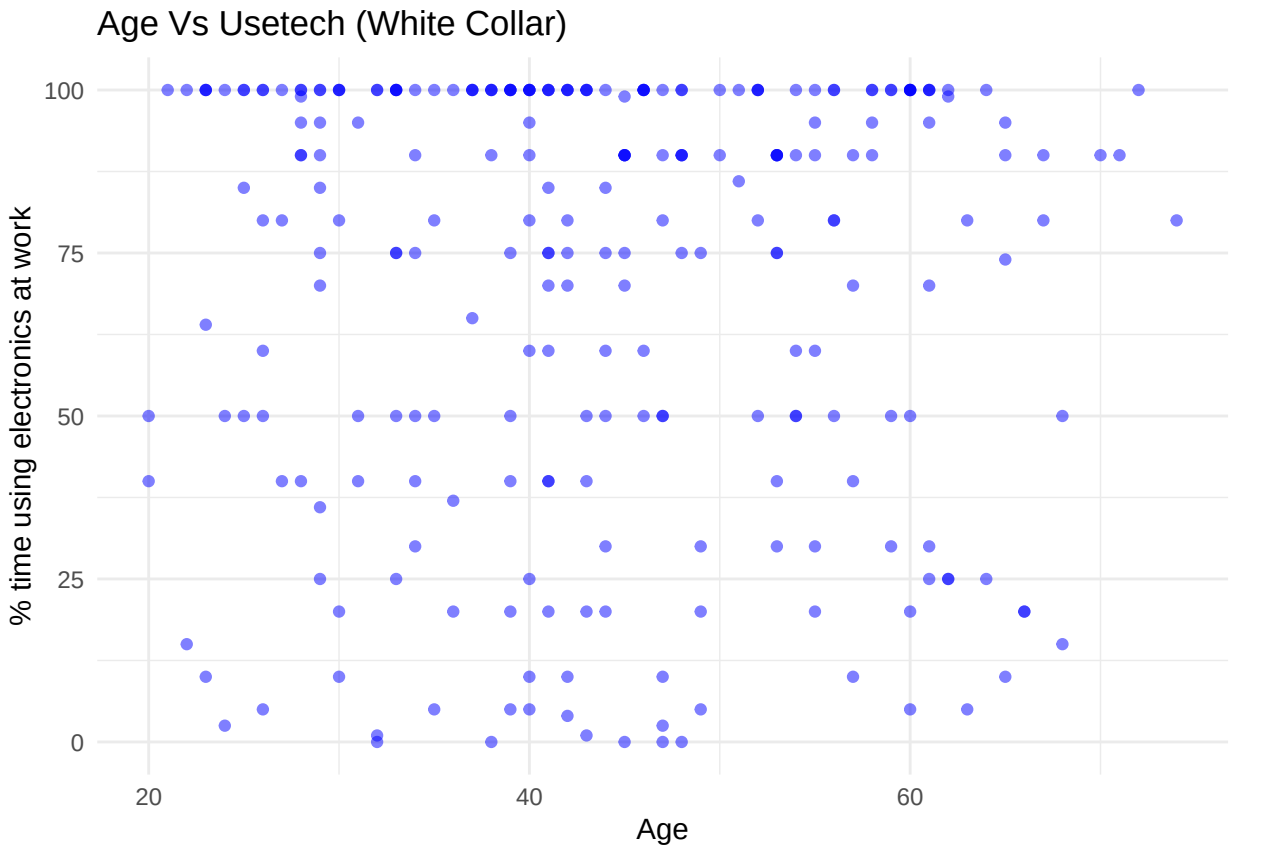
```
                                    labs(x="Age", y="% time using electronics at work", title = "Ag
                                    theme_minimal()

age_vs_usetech_service_scatter <- ds_exp %>%
                                  filter(job_ctg=="Service") %>%
                                  ggplot()+
                                  aes(x=age, y=usetech)+
                                  geom_point(color="blue", alpha=0.5)+
                                  labs(x="Age", y="% time using electronics at work", title = "Age Vs
                                  theme_minimal()

age_vs_usetech_scatter
```



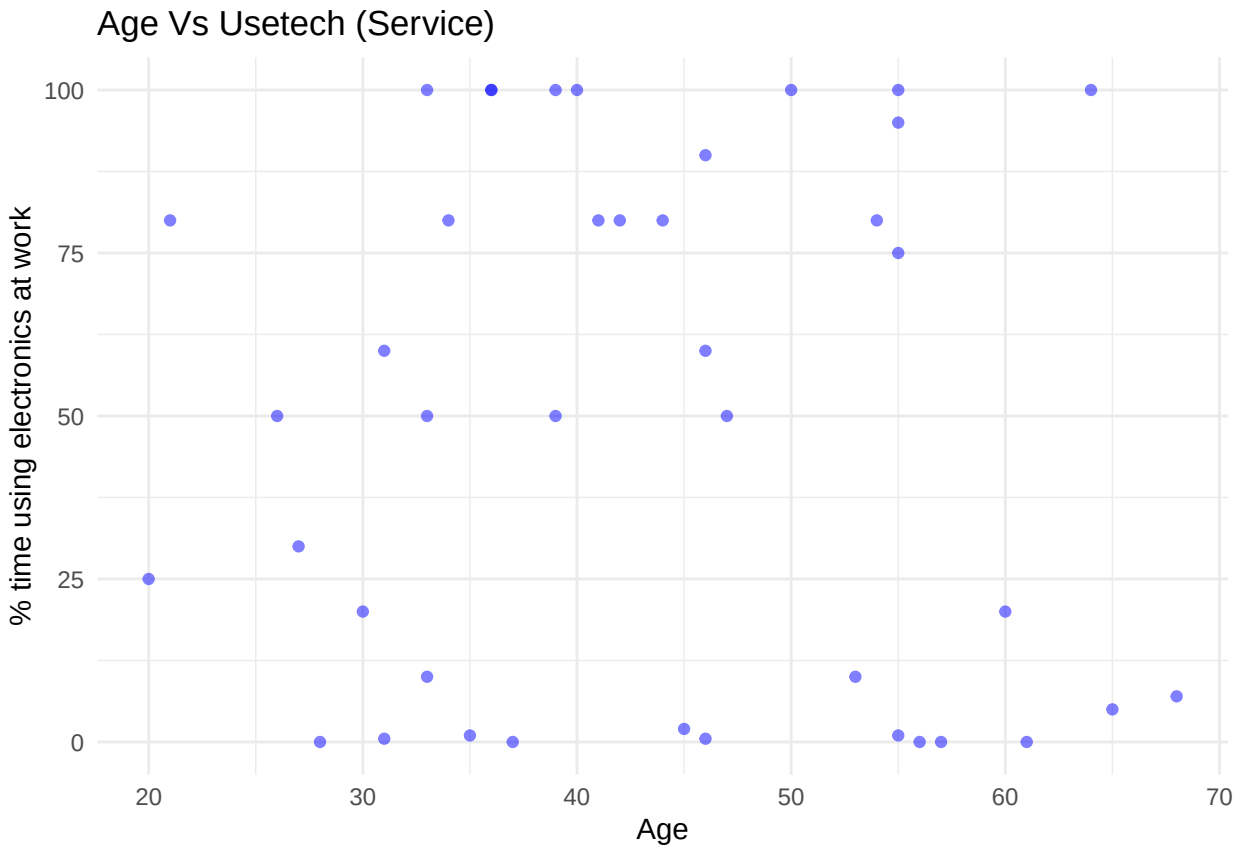

```
age_vs_usetech_bluecollar_scatter
```




```
age_vs_usetech_whitecollar_scatter
```




```
age_vs_usetech_service_scatter
```

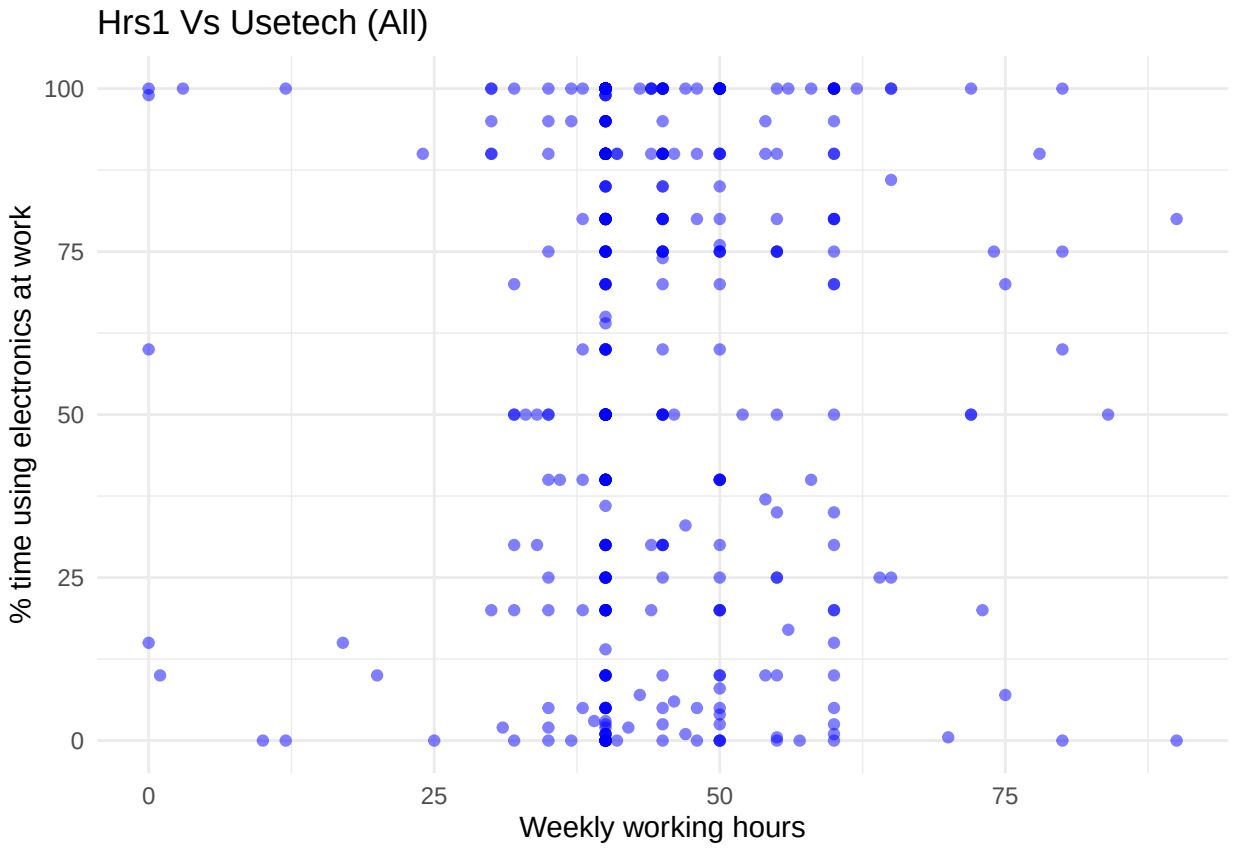



```
## Save the plots.
ggsave("~/lab-2-team-no-l-s/notebooks/plots/age_vs_usetech_scatter.png", plot = age_vs_usetech_scatter,
ggsave("~/lab-2-team-no-l-s/notebooks/plots/age_vs_usetech_bluecollar_scatter.png", plot = age_vs_usete
ggsave("~/lab-2-team-no-l-s/notebooks/plots/age_vs_usetech_whitecollar_scatter.png", plot = age_vs_uset
ggsave("~/lab-2-team-no-l-s/notebooks/plots/age_vs_usetech_service_scatter.png", plot = age_vs_usetech_s
```

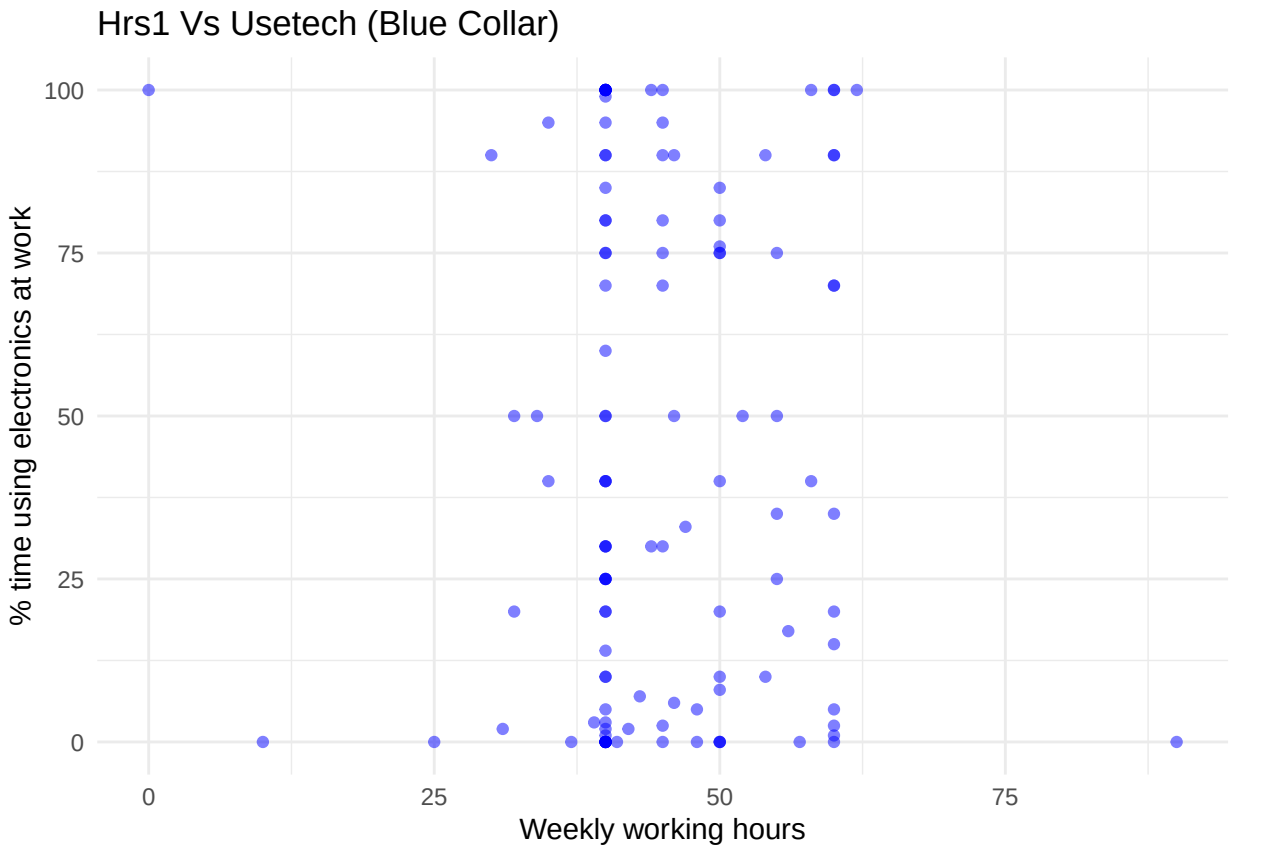
Observation: Shows that there's no proper relationship between percentage of time spent on electronics at work Vs Age.

```
hrs1_vs_usetech_scatter <- ds_exp %>%
                           ggplot()+
                           aes(x=hrs1, y=usetech)+
                           geom_point(color="blue", alpha=0.5)+
                           labs(x="Weekly working hours", y="% time using electronics at work", title =
                           theme_minimal()

hrs1_vs_usetech_bluecollar_scatter <- ds_exp %>%
                                      filter(job_ctg=="Blue Collar") %>%
                                      ggplot()+
                                      aes(x=hrs1, y=usetech)+
                                      geom_point(color="blue", alpha=0.5)+
                                      labs(x="Weekly working hours", y="% time using electronics at w
                                      theme_minimal()

hrs1_vs_usetech_whitecollar_scatter <- ds_exp %>%
                                      filter(job_ctg=="White Collar") %>%
                                      ggplot()+
                                      aes(x=hrs1, y=usetech)+
```

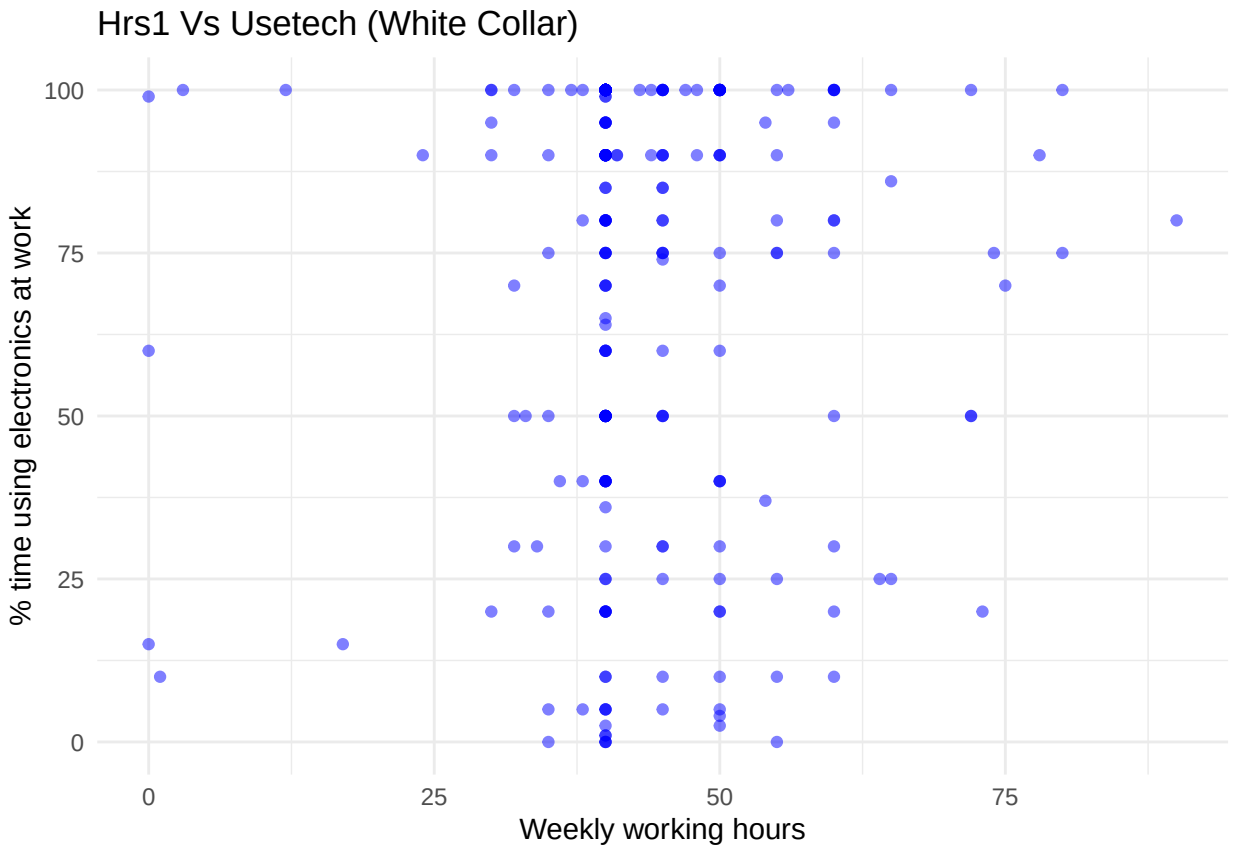
```
                                    geom_point(color="blue", alpha=0.5)+
                                    labs(x="Weekly working hours", y="% time using electronics at w
                                    theme_minimal()

hrs1_vs_usetech_service_scatter <- ds_exp %>%
                                   filter(job_ctg=="Service") %>%
                                   ggplot()+
                                   aes(x=hrs1, y=usetech)+
                                   geom_point(color="blue", alpha=0.5)+
                                   labs(x="Weekly working hours", y="% time using electronics at work"
                                   theme_minimal()

hrs1_vs_usetech_scatter
```



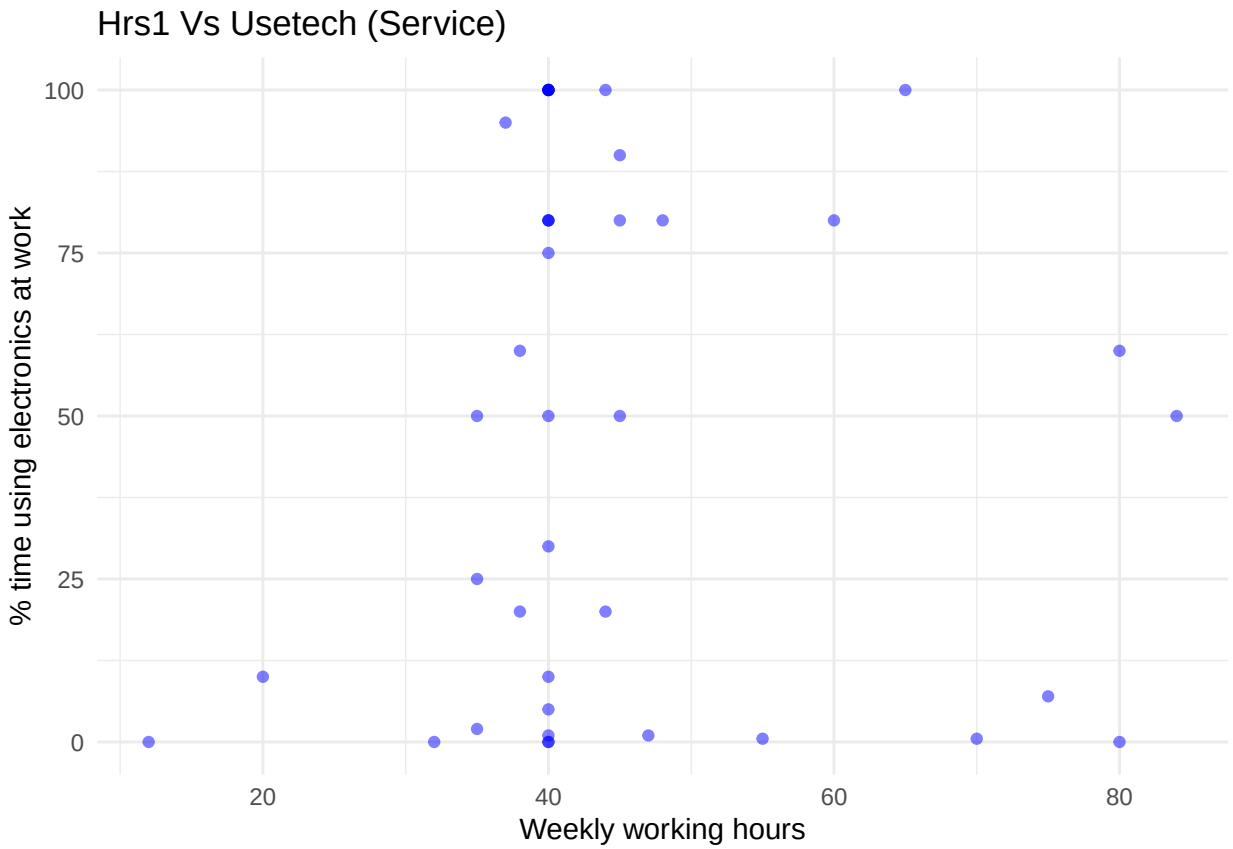

```
hrs1_vs_usetech_bluecollar_scatter
```




```
hrs1_vs_usetech_whitecollar_scatter
```




```
hrs1_vs_usetech_service_scatter
```




```
## Save the plots.
ggsave("~/lab-2-team-no-l-s/notebooks/plots/hrs1_vs_usetech_scatter.png", plot = hrs1_vs_usetech_scatte
ggsave("~/lab-2-team-no-l-s/notebooks/plots/hrs1_vs_usetech_bluecollar_scatter.png", plot = hrs1_vs_use
ggsave("~/lab-2-team-no-l-s/notebooks/plots/hrs1_vs_usetech_whitecollar_scatter.png", plot = hrs1_vs_us
ggsave("~/lab-2-team-no-l-s/notebooks/plots/hrs1_vs_usetech_service_scatter.png", plot = hrs1_vs_usetech
```

Observation: Shows that there's no proper relationship between percentage of time spent on electronics at work Vs hours of time worked per week.

```
model_baseline <- lm(usetech ~ 1, data = ds_exp)
summary(model_baseline)
```

```
##
## Call:
## lm(formula = usetech ~ 1, data = ds_exp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.95 -34.95  15.05  40.05  40.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.954      1.834   32.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.32 on 413 degrees of freedom
```

Observation:

The average value of usetech across all observations in the sample data is approximately 60.41.

Std. Error = 1.816: The standard error of the mean estimate.

t value = 33.26, p-value < 2e-16: This tells us that the mean value is highly statistically significant.

```
model_full <- lm(usetech ~ age + hrs1 + industry + job_ctg, data = ds_exp)
summary(model_full)
```

```
##
## Call:
## lm(formula = usetech ~ age + hrs1 + industry + job_ctg, data = ds_exp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.98 -30.91  10.14  27.93  67.23
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       34.397342  10.302047
## age                                               -0.072130   0.135618
## hrs1                                               0.007483   0.149800
## industryEducation and Health Services              5.380184   7.194423
## industryFinancial Activities                      17.855399   7.978924
## industryInformation                               25.947162   9.817081
## industryLeisure and Hospitality                   17.593839  10.958234
## industryManufacturing                             21.774861   7.945506
## industryNatural Resources and Mining              10.581703  15.416009
## industryNo Match                                  31.413524  14.757708
## industryOther Services (except Public Administration) 14.286114   9.844194
## industryProfessional and Business Services        26.105549   8.606377
## industryTrade, Transportation, and Utilities      14.615614   6.822314
## job_ctgService                                     2.743686   6.725676
## job_ctgWhite Collar                               22.486428   4.171913
##                                                   t value Pr(>|t|)
## (Intercept)                                         3.339  0.00092 ***
## age                                                -0.532  0.59512
## hrs1                                                0.050  0.96018
## industryEducation and Health Services               0.748  0.45500
## industryFinancial Activities                        2.238  0.02578 *
## industryInformation                                 2.643  0.00854 **
## industryLeisure and Hospitality                     1.606  0.10917
## industryManufacturing                               2.741  0.00641 **
## industryNatural Resources and Mining                0.686  0.49285
## industryNo Match                                    2.129  0.03390 *
## industryOther Services (except Public Administration)  1.451  0.14750
## industryProfessional and Business Services          3.033  0.00258 **
## industryTrade, Transportation, and Utilities        2.142  0.03277 *
## job_ctgService                                      0.408  0.68354
## job_ctgWhite Collar                                 5.390 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.04 on 399 degrees of freedom
```

```
## Multiple R-squared:  0.1481, Adjusted R-squared:  0.1183
## F-statistic: 4.956 on 14 and 399 DF,  p-value: 1.581e-08
```

Observation:

```
anova(model_baseline, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: usetech ~ 1
## Model 2: usetech ~ age + hrs1 + industry + job_ctg
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    413 575071
## 2    399 489879 14     85192 4.9563 1.581e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_1 <- lm(usetech ~ age, data = ds_exp)
model_2 <- lm(usetech ~ age + hrs1, data = ds_exp)
model_3 <- lm(usetech ~ age + hrs1 + industry, data = ds_exp)
anova(model_baseline, model_1, model_2, model_3, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: usetech ~ 1
## Model 2: usetech ~ age
## Model 3: usetech ~ age + hrs1
## Model 4: usetech ~ age + hrs1 + industry
## Model 5: usetech ~ age + hrs1 + industry + job_ctg
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    413 575071
## 2    412 574968  1       102  0.0832 0.7731748
## 3    411 574845  1       123  0.1004 0.7515388
## 4    401 531468 10     43378  3.5331 0.0001716 ***
## 5    399 489879  2     41589 16.9368  8.71e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 1 vs. Model 2 (Adding age) The RSS decreases marginally from 564,112 to 564,082 (Change in RSS = 30), with an F-value of 0.0238 and p-value of 0.877.

Interpretation: Age does not significantly explain variation in usetech. This suggests that tech use is relatively stable across age groups in this sample.

- Model 2 vs. Model 3 (Adding hrs1) The RSS drops to 561,827 (Change in RSS = 2,255), with an F-value of 1.7963 and p-value of 0.1809.

Interpretation: While the model fit improves slightly, the effect of hours worked is not statistically significant. More work hours do not strongly predict higher tech use.

- Model 3 vs. Model 4 (Adding industry) The RSS further decreases to 538,217 (Change in RSS = 23,610), with an F-value of 2.0896 and a statistically significant p-value of 0.0294.

Interpretation: Industry explains a meaningful amount of variance in tech use. This indicates that technology use is influenced by sector-level practices and work environments.

- Model 4 vs. Model 5 (Adding job_ctg) The final RSS drops to 502,176 (Change in RSS = 36,041), with an F-value of 14.3539 and a highly significant p-value of 9.54e-07.

Interpretation: Job category significantly enhances the model. Even within the same industry, the nature of a person's role determines their likelihood of using technology at work.

=> This stepwise model comparison highlights that:

Demographic and individual-level predictors (like age and hrs1) do not significantly explain technology use.

Workplace context — including industry and job type — plays a more substantial role.

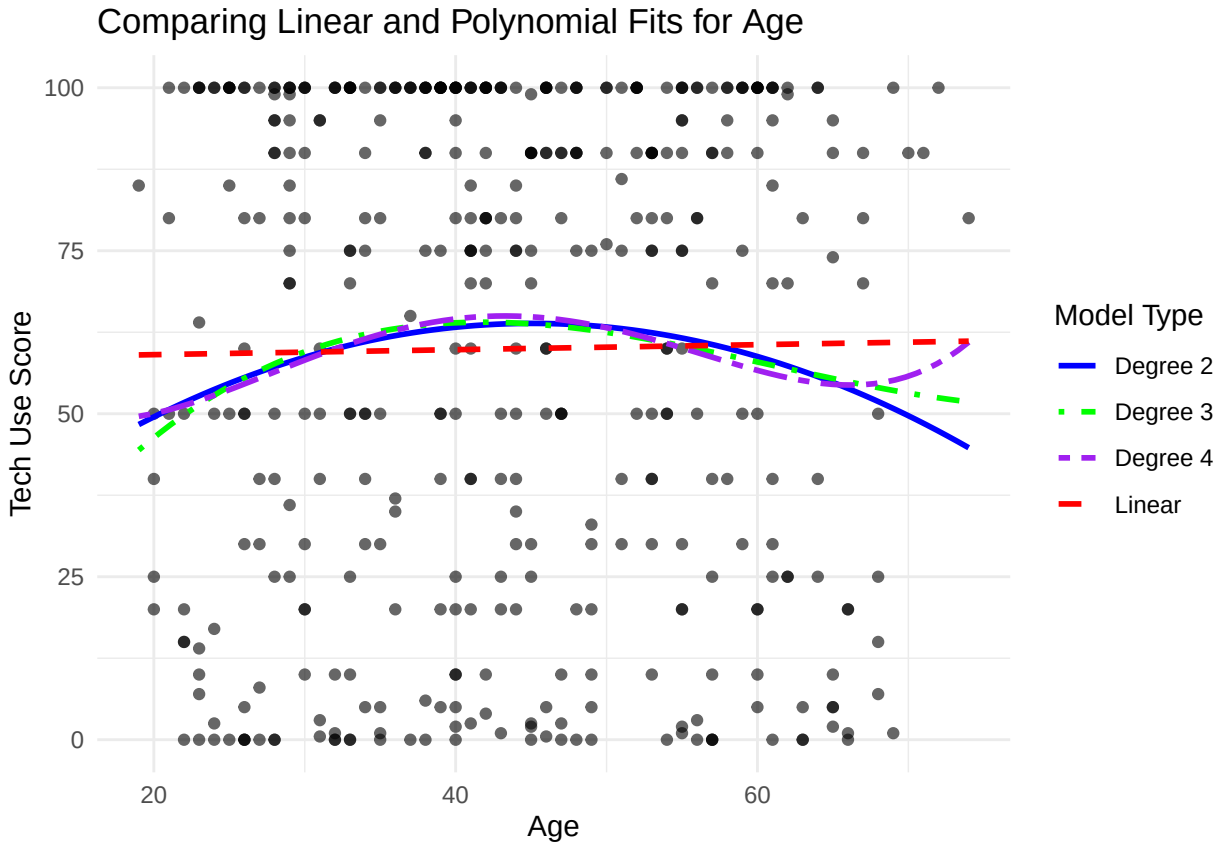
These results underscore the importance of organizational structures and job roles in shaping how individuals engage with technology in professional settings.

«««< Updated upstream

And now, to show transformations with polynomials, and what a graph looks like when expanding age into multidegree polynomials.

```
ggplot(ds_exp, aes(x = age, y = usetech)) +
  geom_point(alpha = 0.6) +

  stat_smooth(
    mapping = aes(color = "Degree 2"),
    method = "lm",
    formula = y ~ poly(x, 2),
    se = FALSE
  ) +

  stat_smooth(
    mapping = aes(color = "Degree 3"),
    method = "lm",
    formula = y ~ poly(x, 3),
    linetype = "dotdash",
    se = FALSE
  ) +

  stat_smooth(
    mapping = aes(color = "Degree 4"),
    method = "lm",
    formula = y ~ poly(x, 4),
    linetype = "twodash",
    se = FALSE
  ) +

  stat_smooth(
    mapping = aes(color = "Linear"),
    method = "lm",
    formula = y ~ x,
    linetype = "dashed",
    se = FALSE
  ) +

  scale_color_manual(
    name = "Model Type",
    values = c(
      "Linear" = "red",
      "Degree 2" = "blue",
      "Degree 3" = "green",
```

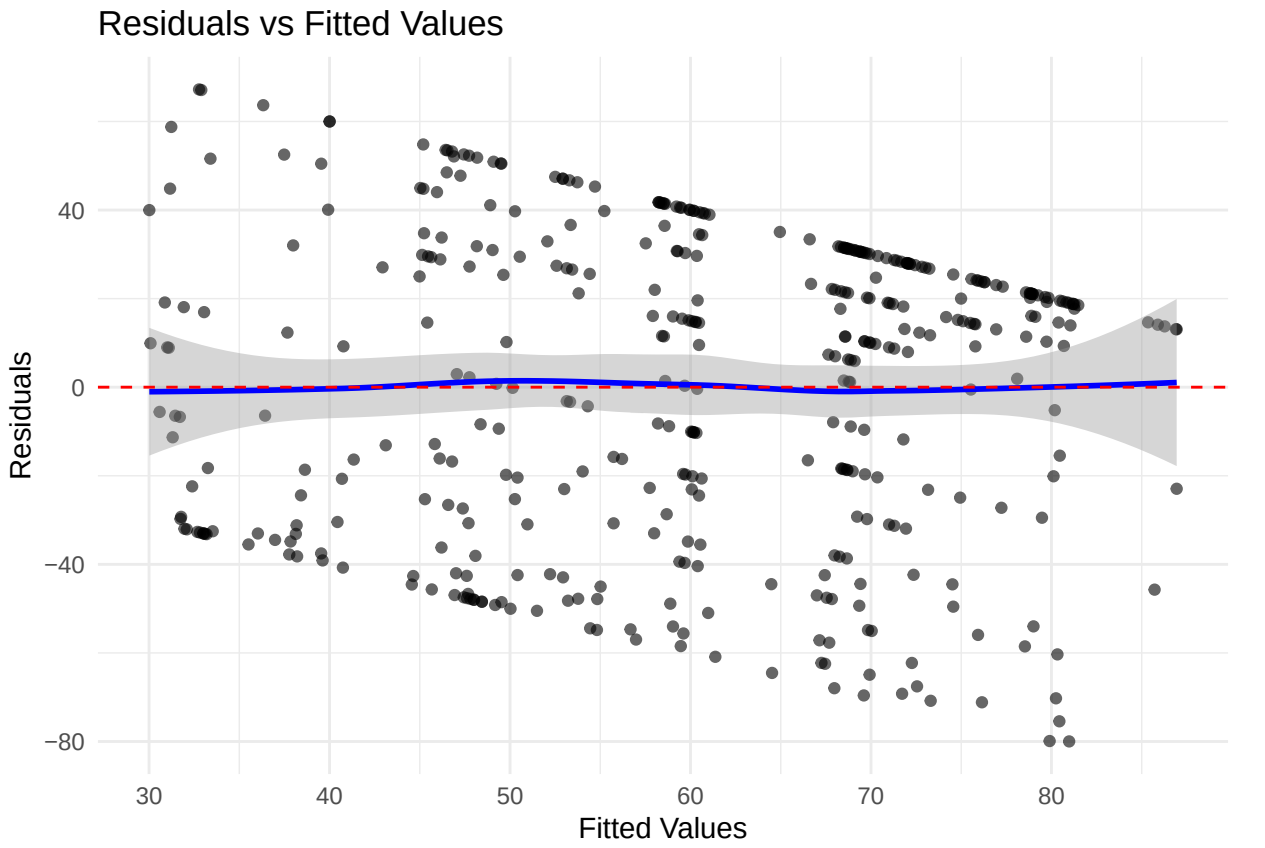
```
      "Degree 4" = "purple"
    )
  ) +

  labs(
    title = "Comparing Linear and Polynomial Fits for Age",
    x = "Age",
    y = "Tech Use Score",
    color = "Model Type" # This ensures the legend title is correct
  ) +
  theme_minimal() +
  theme(legend.text = element_text(color = names(scale_color_manual()$palette)))
```




```
plot_data <- data.frame(
  Fitted = fitted(model_full),
  Residuals = resid(model_full)
)

# Make the plot
ggplot(plot_data, aes(x = Fitted, y = Residuals)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = TRUE, color = "blue", linewidth = 1) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
```

```
       y = "Residuals") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Below is using stargazer to show the comparison between the results of lm(usetech ~ age + hrs1 + industry) and lm(usetech ~ age + hrs1 + industry + job_ctg).

```
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
model1 <- lm(usetech ~ age + hrs1 + industry, data = ds_exp)
model2 <- lm(usetech ~ age + hrs1 + industry + job_ctg, data = ds_exp)

stargazer(model1, model2,
          type = "text",       # Use "html" or "latex" for formatted output
          title = "Comparison of Regression Models",
          column.labels = c("Without Job Category", "With Job Category"),
          dep.var.labels = "Technology Use (usetech)",
          align = TRUE,
          no.space = TRUE)
```

```
##
## Comparison of Regression Models
```

```
## ==========================================================================================
##                                                              Dependent variable:
##                                                  ---------------------------------------
##                                                          Technology Use (usetech)
##                                                   Without Job Category     With Job Category
##                                                            (1)                      (2)
## ------------------------------------------------------------------------------------------
## age                                                      -0.036                   -0.072
##                                                          (0.141)                  (0.136)
## hrs1                                                     -0.031                    0.007
##                                                          (0.155)                  (0.150)
## industryEducation and Health Services                   15.555**                   5.380
##                                                          (7.082)                  (7.194)
## industryFinancial Activities                            28.137***                17.855**
##                                                          (8.052)                  (7.979)
## industryInformation                                     32.363***                25.947***
##                                                          (9.754)                  (9.817)
## industryLeisure and Hospitality                          21.664*                  17.594
##                                                         (11.346)                 (10.958)
## industryManufacturing                                   29.282***                21.775***
##                                                          (8.116)                  (7.946)
## industryNatural Resources and Mining                     15.689                   10.582
##                                                         (15.990)                 (15.416)
## industryNo Match                                        48.394***                31.414**
##                                                         (14.992)                 (14.758)
## industryOther Services (except Public Administration)    19.147*                  14.286
##                                                         (10.018)                  (9.844)
## industryProfessional and Business Services              39.125***                26.106***
##                                                          (8.600)                  (8.606)
## industryTrade, Transportation, and Utilities            22.828***                14.616**
##                                                          (6.841)                  (6.822)
## job_ctgService                                                                     2.744
##                                                                                   (6.726)
## job_ctgWhite Collar                                                              22.486***
##                                                                                   (4.172)
## Constant                                                40.439***                34.397***
##                                                         (10.632)                 (10.302)
## ------------------------------------------------------------------------------------------
## Observations                                              414                      414
## R2                                                       0.076                    0.148
## Adjusted R2                                              0.048                    0.118
## Residual Std. Error                                36.405 (df = 401)        35.039 (df = 399)
## F Statistic                                     2.742*** (df = 12; 401) 4.956*** (df = 14; 399
## ==========================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.0
```