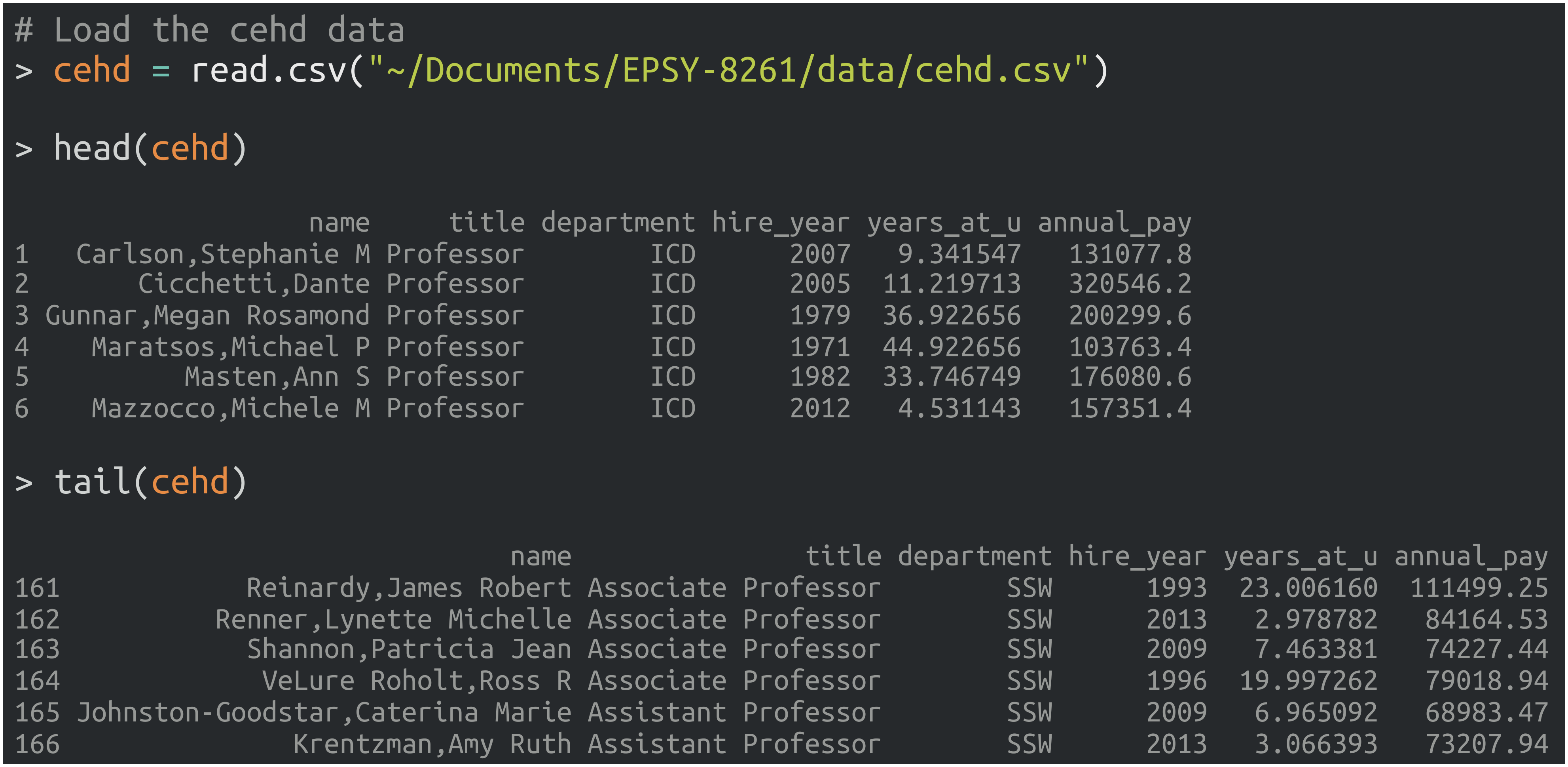
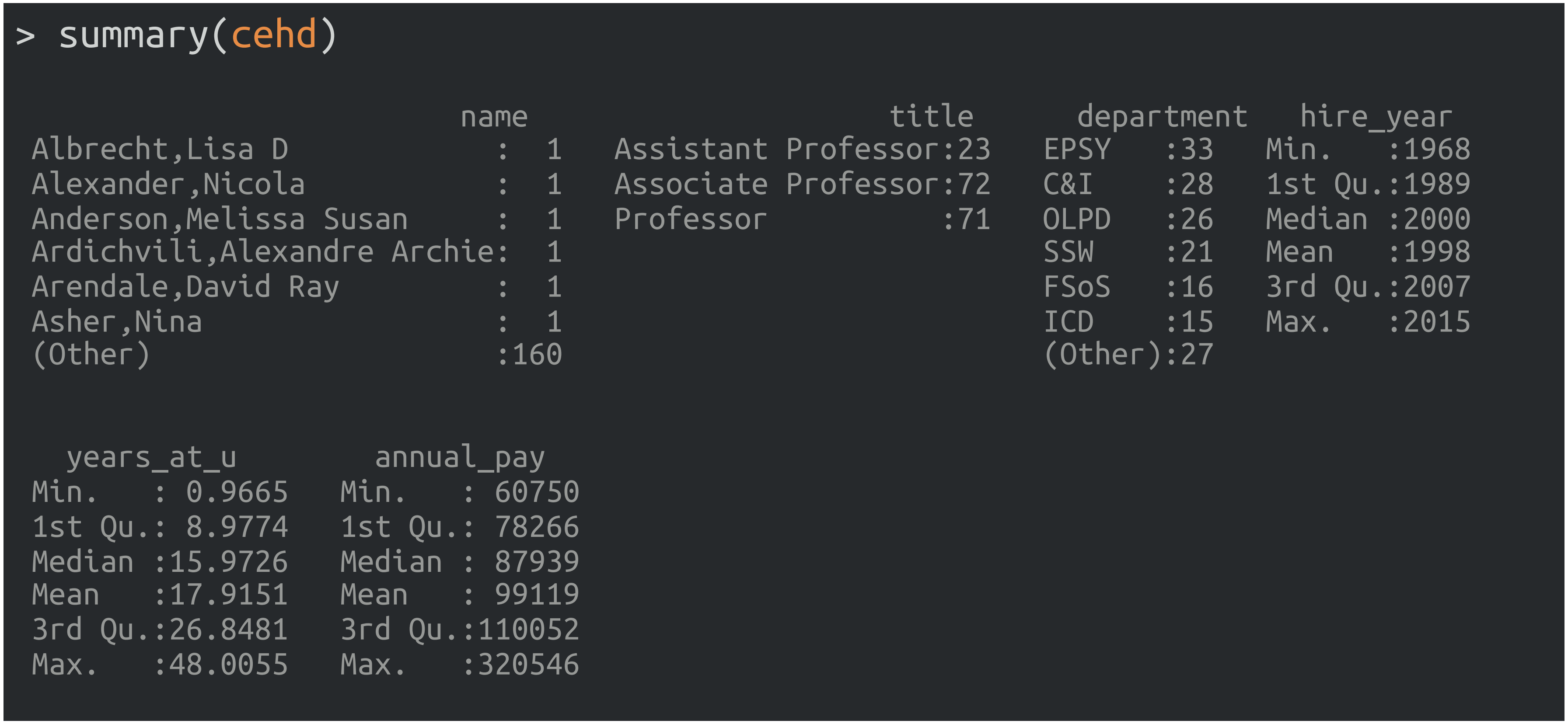
Data Wrangling with dplyr

Andrew Zieffler



```
# Load the cehd data
> cehd = read.csv("~/Documents/EPSY-8261/data/cehd.csv")

> head(cehd)

                   name     title department hire_year years_at_u annual_pay
1   Carlson,Stephanie M Professor        ICD      2007   9.341547   131077.8
2       Cicchetti,Dante Professor        ICD      2005  11.219713   320546.2
3 Gunnar,Megan Rosamond Professor        ICD      1979  36.922656   200299.6
4     Maratsos,Michael P Professor        ICD      1971  44.922656   103763.4
5           Masten,Ann S Professor        ICD      1982  33.746749   176080.6
6     Mazzocco,Michele M Professor        ICD      2012   4.531143   157351.4

> tail(cehd)

                                name               title department hire_year years_at_u annual_pay
161            Reinardy,James Robert Associate Professor        SSW      1993  23.006160  111499.25
162          Renner,Lynette Michelle Associate Professor        SSW      2013   2.978782   84164.53
163            Shannon,Patricia Jean Associate Professor        SSW      2009   7.463381   74227.44
164             VeLure Roholt,Ross R Associate Professor        SSW      1996  19.997262   79018.94
165 Johnston-Goodstar,Caterina Marie Assistant Professor        SSW      2009   6.965092   68983.47
166               Krentzman,Amy Ruth Assistant Professor        SSW      2013   3.066393   73207.94
```

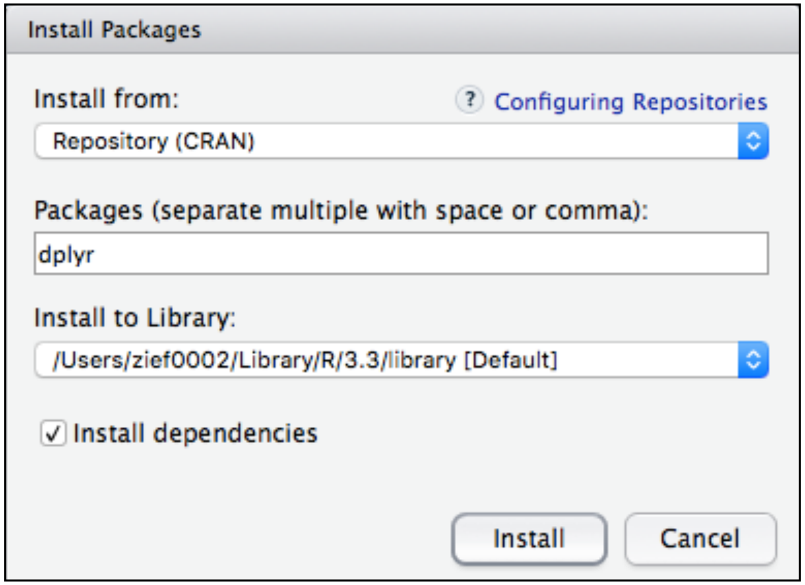


```
> summary(cehd)

                          name                     title      department   hire_year
 Albrecht,Lisa D            :  1   Assistant Professor:23   EPSY   :33   Min.   :1968
 Alexander,Nicola           :  1   Associate Professor:72   C&I    :28   1st Qu.:1989
 Anderson,Melissa Susan     :  1   Professor          :71   OLPD   :26   Median :2000
 Ardichvili,Alexandre Archie:  1                            SSW    :21   Mean   :1998
 Arendale,David Ray         :  1                            FSoS   :16   3rd Qu.:2007
 Asher,Nina                 :  1                            ICD    :15   Max.   :2015
 (Other)                    :160                            (Other):27


   years_at_u        annual_pay
 Min.   : 0.9665   Min.   : 60750
 1st Qu.: 8.9774   1st Qu.: 78266
 Median :15.9726   Median : 87939
 Mean   :17.9151   Mean   : 99119
 3rd Qu.:26.8481   3rd Qu.:110052
 Max.   :48.0055   Max.   :320546
```

Install and load the **dplyr** Package

Using the RStudio GUI…

- Click the **Packages** tab.
- Click **Install Packages**.
- Enter *dplyr* in the text box.
- Click **Install**.



Install Packages

Install from: Repository (CRAN)

Packages (separate multiple with space or comma): dplyr

Install to Library: /Users/zief0002/Library/R/3.3/library [Default]

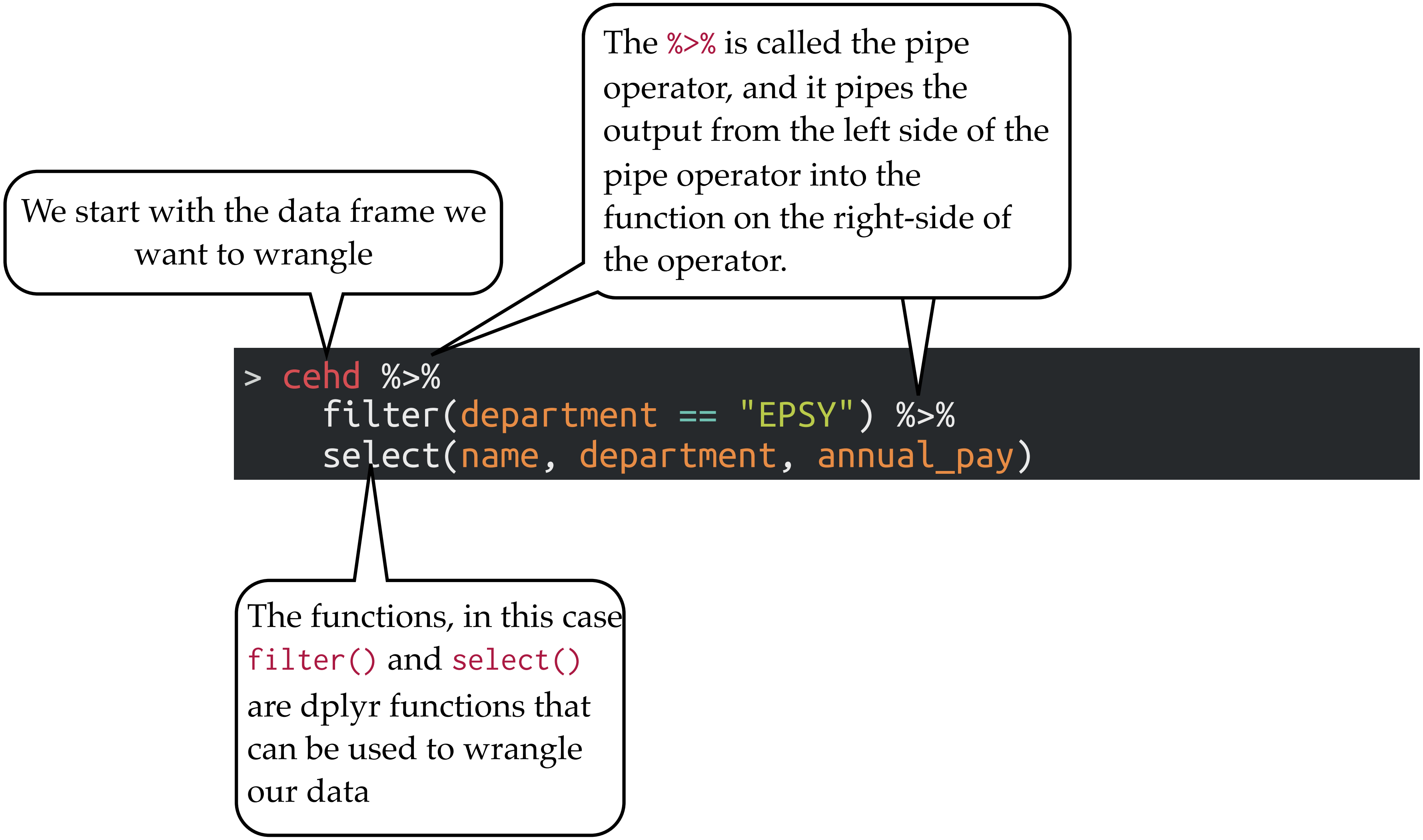
☑ Install dependencies

…or directly from the R command line…

```
> install.packages("dplyr", dependencies = TRUE)
```

```
# After installing the package, load the dplyr library
> library(dplyr)
```

Understanding the Basic Syntax



We start with the data frame we want to wrangle

The %>% is called the pipe operator, and it pipes the output from the left side of the pipe operator into the function on the right-side of the operator.

```
> cehd %>%
    filter(department == "EPSY") %>%
    select(name, department, annual_pay)
```

The functions, in this case `filter()` and `select()` are dplyr functions that can be used to wrangle our data

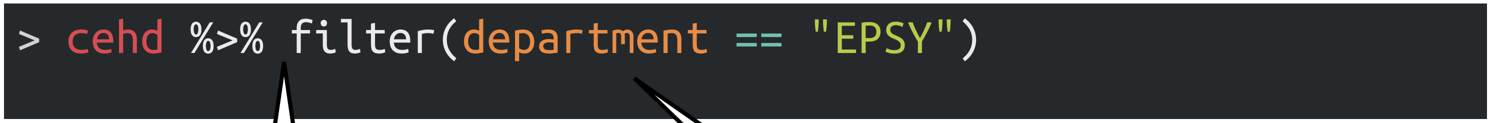
Common dplyr Functions

Here are some common operations that we want to use with data to "wrangle" it into shape and the corresponding **dplyr** functions.

- Select a subset of rows from a data frame. Use the `filter()` function.
- Select a subset of columns from a data frame. Use the `select()` function.
- Add new columns that are functions of existing columns. Use the `mutate()` function.
- Sort and re-order data in a data frame. Use the `arrange()` function.
- Compute summaries of a data frame. Use the `summarize()` function.
- Group the data to carry out computations for each group. Use the `group_by()` function.

Select a Subset of Rows

To select a subset of rows, we will use the `filter()` function. The argument(s) for this function are expressions that filter the data frame.

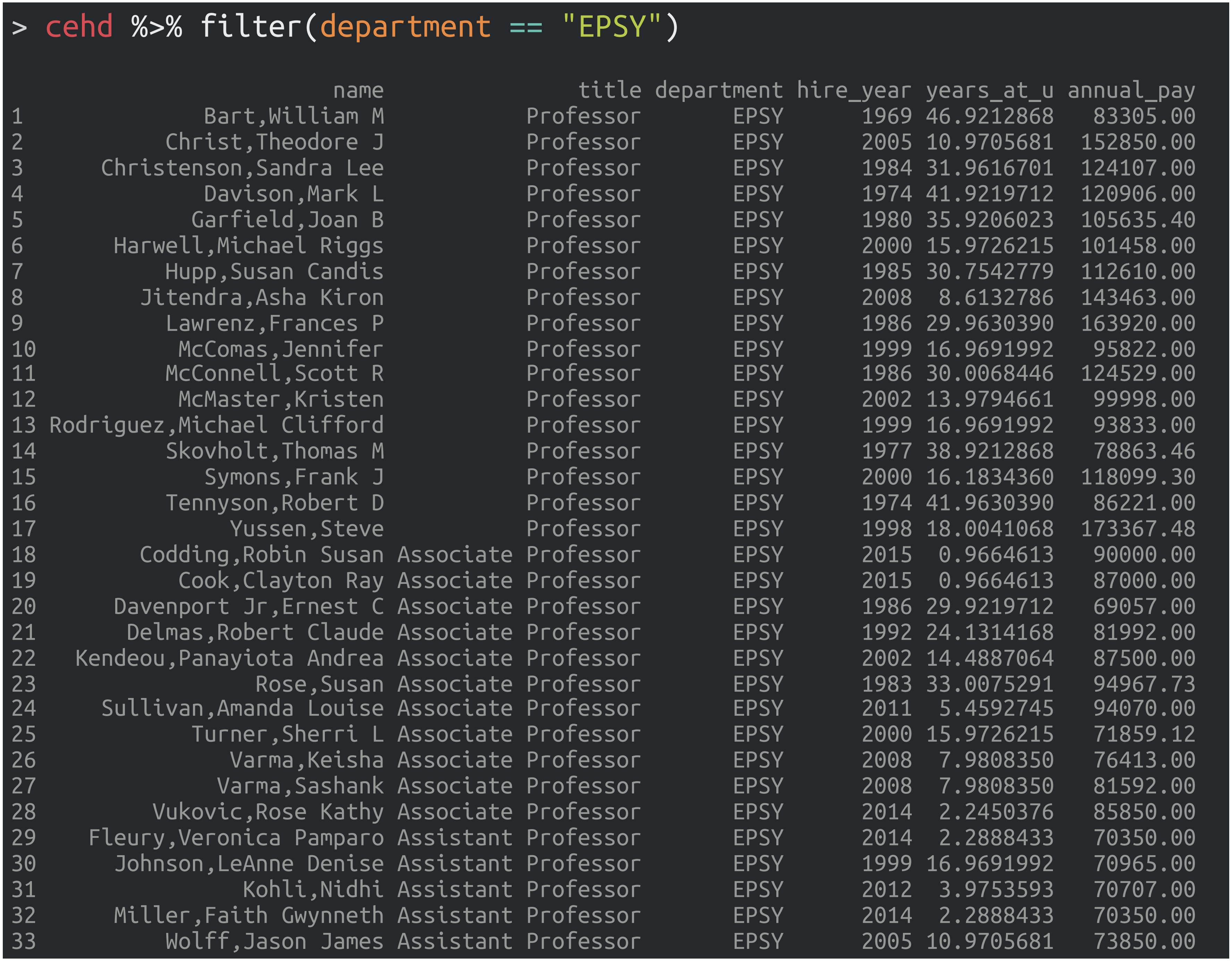


```
> cehd %>% filter(department == "EPSY")
```

The `data=` argument indicates the source data frame.

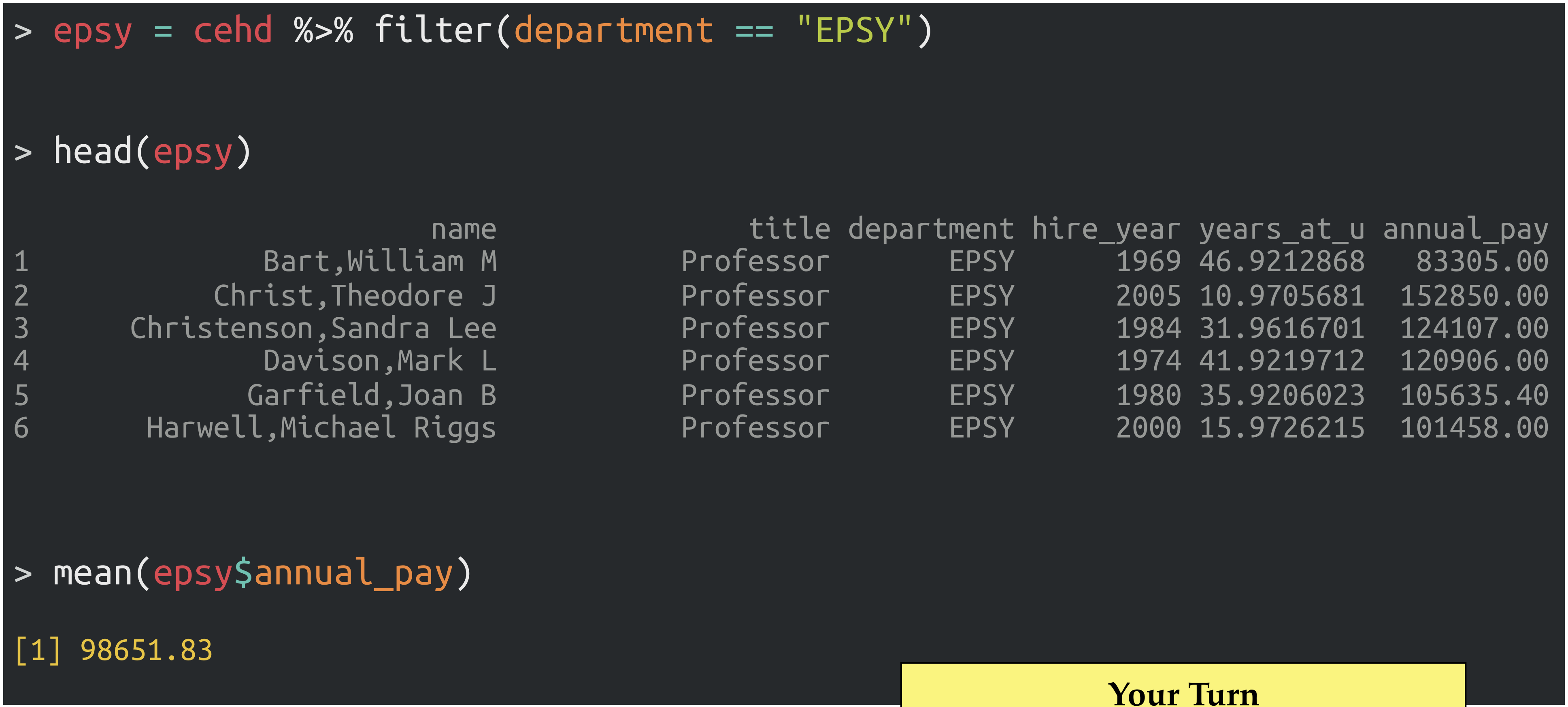
The `aes=` argument sets the aesthetic mapping(s).

Here we are selecting only the rows where the department **is equal to** (==) the string “EPSY”. In other words, we are selecting the faculty from the Educational Psychology department.



```
> cehd %>% filter(department == "EPSY")

                         name               title department hire_year years_at_u annual_pay
1              Bart,William M           Professor       EPSY      1969 46.9212868   83305.00
2           Christ,Theodore J           Professor       EPSY      2005 10.9705681  152850.00
3      Christenson,Sandra Lee           Professor       EPSY      1984 31.9616701  124107.00
4              Davison,Mark L           Professor       EPSY      1974 41.9219712  120906.00
5             Garfield,Joan B           Professor       EPSY      1980 35.9206023  105635.40
6       Harwell,Michael Riggs           Professor       EPSY      2000 15.9726215  101458.00
7           Hupp,Susan Candis           Professor       EPSY      1985 30.7542779  112610.00
8         Jitendra,Asha Kiron           Professor       EPSY      2008  8.6132786  143463.00
9           Lawrenz,Frances P           Professor       EPSY      1986 29.9630390  163920.00
10           McComas,Jennifer           Professor       EPSY      1999 16.9691992   95822.00
11          McConnell,Scott R           Professor       EPSY      1986 30.0068446  124529.00
12           McMaster,Kristen           Professor       EPSY      2002 13.9794661   99998.00
13 Rodriguez,Michael Clifford           Professor       EPSY      1999 16.9691992   93833.00
14          Skovholt,Thomas M           Professor       EPSY      1977 38.9212868   78863.46
15             Symons,Frank J           Professor       EPSY      2000 16.1834360  118099.30
16          Tennyson,Robert D           Professor       EPSY      1974 41.9630390   86221.00
17               Yussen,Steve           Professor       EPSY      1998 18.0041068  173367.48
18        Codding,Robin Susan Associate Professor       EPSY      2015  0.9664613   90000.00
19           Cook,Clayton Ray Associate Professor       EPSY      2015  0.9664613   87000.00
20      Davenport Jr,Ernest C Associate Professor       EPSY      1986 29.9219712   69057.00
21       Delmas,Robert Claude Associate Professor       EPSY      1992 24.1314168   81992.00
22   Kendeou,Panayiota Andrea Associate Professor       EPSY      2002 14.4887064   87500.00
23                 Rose,Susan Associate Professor       EPSY      1983 33.0075291   94967.73
24     Sullivan,Amanda Louise Associate Professor       EPSY      2011  5.4592745   94070.00
25            Turner,Sherri L Associate Professor       EPSY      2000 15.9726215   71859.12
26               Varma,Keisha Associate Professor       EPSY      2008  7.9808350   76413.00
27              Varma,Sashank Associate Professor       EPSY      2008  7.9808350   81592.00
28         Vukovic,Rose Kathy Associate Professor       EPSY      2014  2.2450376   85850.00
29    Fleury,Veronica Pamparo Assistant Professor       EPSY      2014  2.2888433   70350.00
30      Johnson,LeAnne Denise Assistant Professor       EPSY      1999 16.9691992   70965.00
31                Kohli,Nidhi Assistant Professor       EPSY      2012  3.9753593   70707.00
32      Miller,Faith Gwynneth Assistant Professor       EPSY      2014  2.2888433   70350.00
33          Wolff,Jason James Assistant Professor       EPSY      2005 10.9705681   73850.00
```

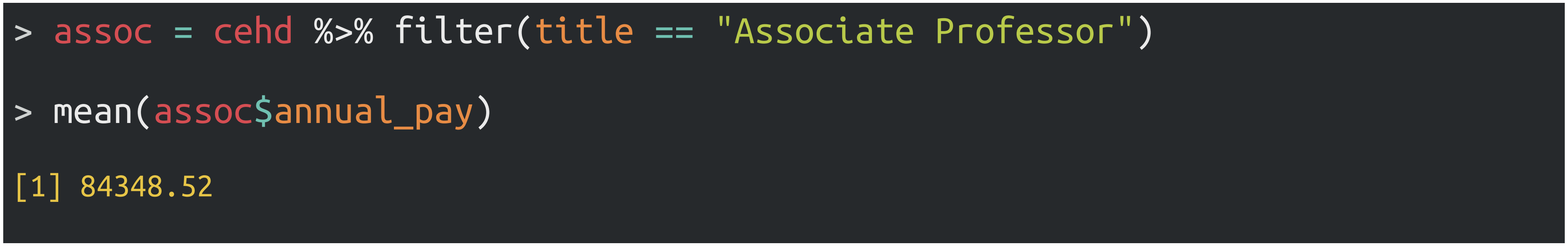
Note the output is just printed to the screen. If you want to keep the filtered data or operate on it further, you need to rite the output into an object.



```
> epsy = cehd %>% filter(department == "EPSY")


> head(epsy)

                         name               title department hire_year years_at_u annual_pay
1              Bart,William M           Professor       EPSY      1969 46.9212868   83305.00
2           Christ,Theodore J           Professor       EPSY      2005 10.9705681  152850.00
3      Christenson,Sandra Lee           Professor       EPSY      1984 31.9616701  124107.00
4              Davison,Mark L           Professor       EPSY      1974 41.9219712  120906.00
5             Garfield,Joan B           Professor       EPSY      1980 35.9206023  105635.40
6       Harwell,Michael Riggs           Professor       EPSY      2000 15.9726215  101458.00


> mean(epsy$annual_pay)

[1] 98651.83
```

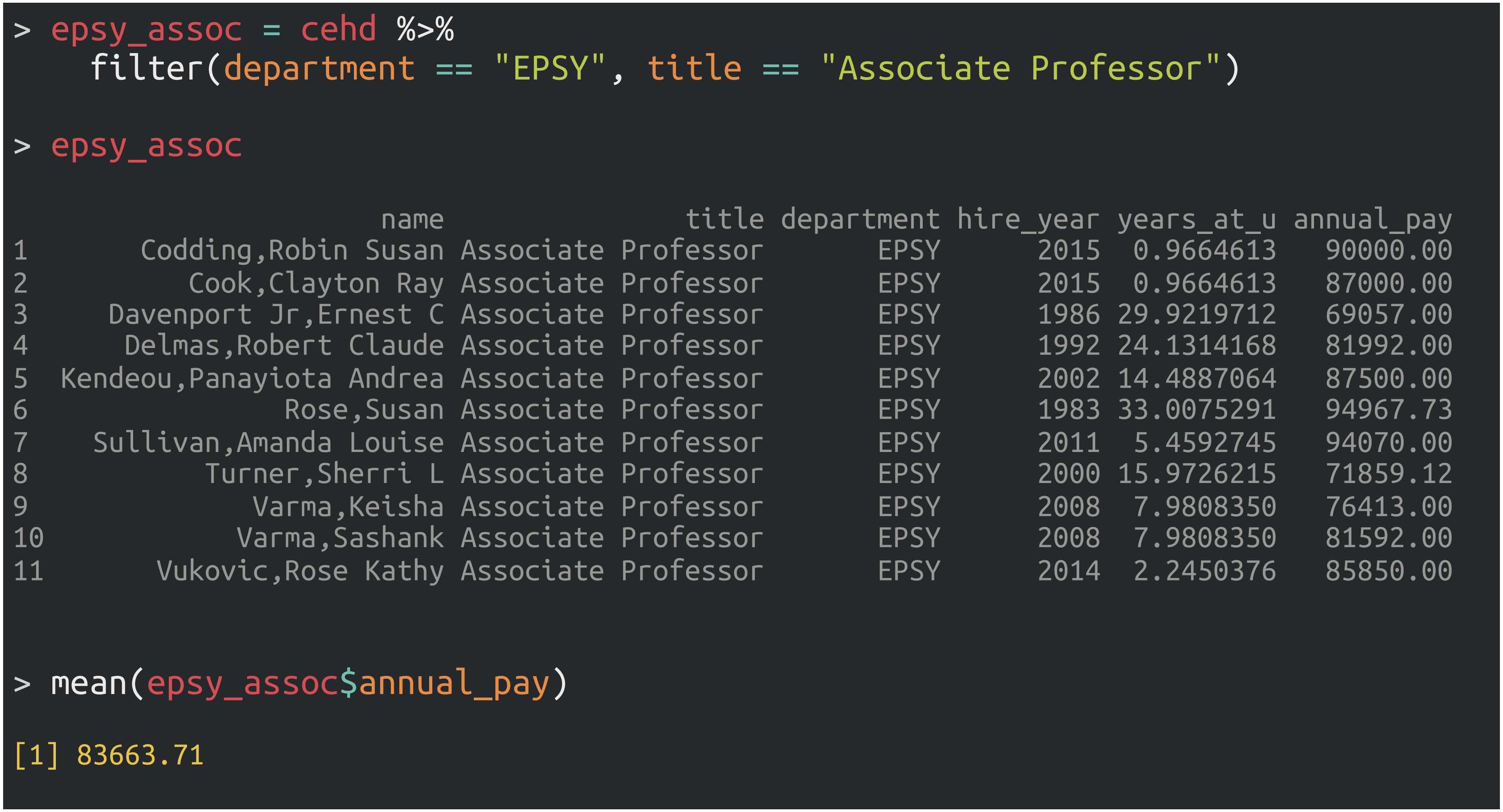
Your Turn

Find the average annual pay for Associate Professors.



```
> assoc = cehd %>% filter(title == "Associate Professor")

> mean(assoc$annual_pay)

[1] 84348.52
```

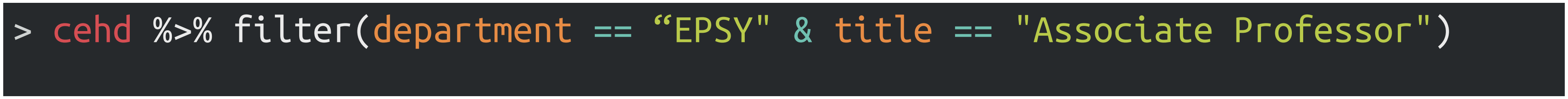
You can filter on multiple attributes by adding additional arguments.



```
> epsy_assoc = cehd %>%
    filter(department == "EPSY", title == "Associate Professor")

> epsy_assoc

                       name               title department hire_year years_at_u annual_pay
1       Codding,Robin Susan Associate Professor       EPSY      2015  0.9664613   90000.00
2          Cook,Clayton Ray Associate Professor       EPSY      2015  0.9664613   87000.00
3     Davenport Jr,Ernest C Associate Professor       EPSY      1986 29.9219712   69057.00
4      Delmas,Robert Claude Associate Professor       EPSY      1992 24.1314168   81992.00
5  Kendeou,Panayiota Andrea Associate Professor       EPSY      2002 14.4887064   87500.00
6                Rose,Susan Associate Professor       EPSY      1983 33.0075291   94967.73
7    Sullivan,Amanda Louise Associate Professor       EPSY      2011  5.4592745   94070.00
8           Turner,Sherri L Associate Professor       EPSY      2000 15.9726215   71859.12
9              Varma,Keisha Associate Professor       EPSY      2008  7.9808350   76413.00
10            Varma,Sashank Associate Professor       EPSY      2008  7.9808350   81592.00
11       Vukovic,Rose Kathy Associate Professor       EPSY      2014  2.2450376   85850.00


> mean(epsy_assoc$annual_pay)

[1] 83663.71
```

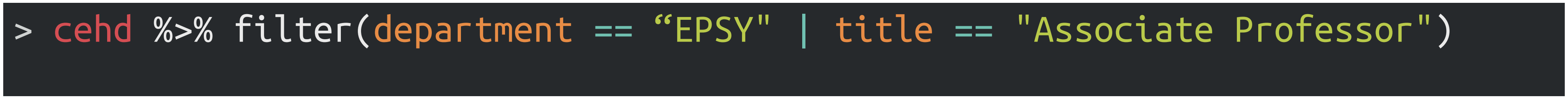
Linking Expressions: AND and OR

When we include multiple expressions in the `filter()` function, they are linked using the AND (&) operator. This means that both expressions have to evaluate as TRUE to be included.

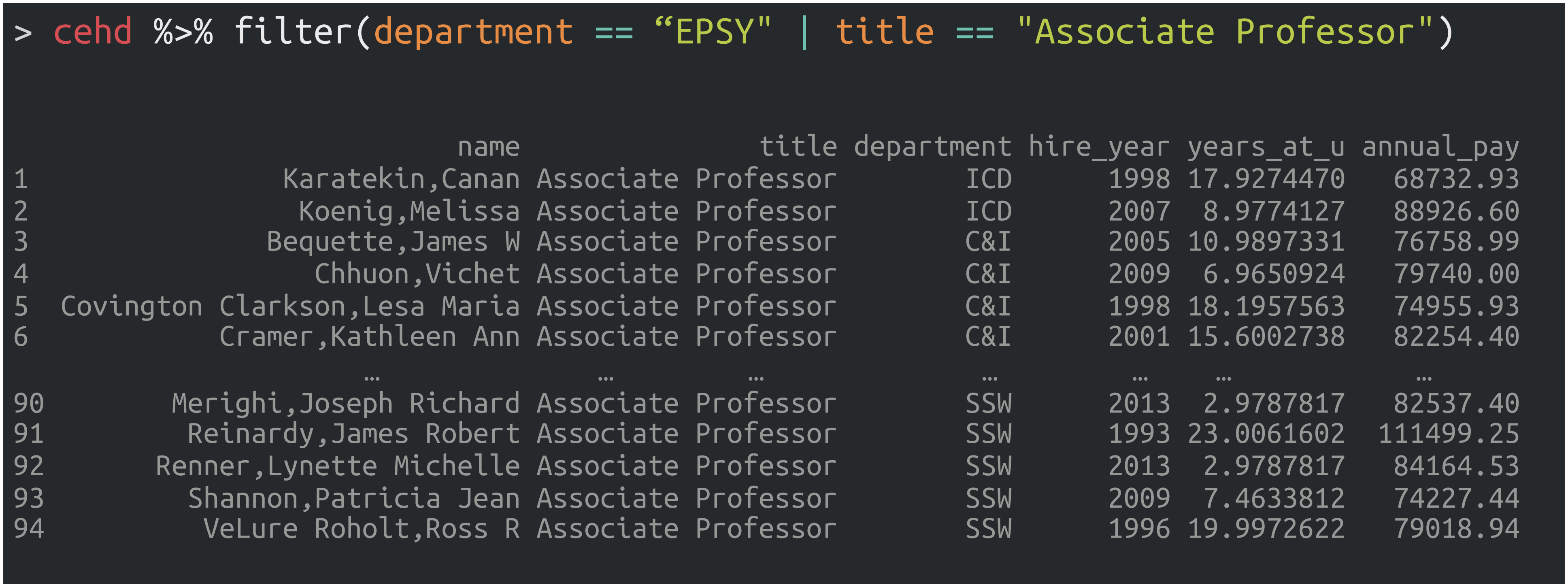


```
> cehd %>% filter(department == “EPSY" & title == "Associate Professor")
```

We can also `filter()` using the OR (|) operator. This means that if either expression evaluates as TRUE it is included.



```
> cehd %>% filter(department == “EPSY" | title == "Associate Professor")
```



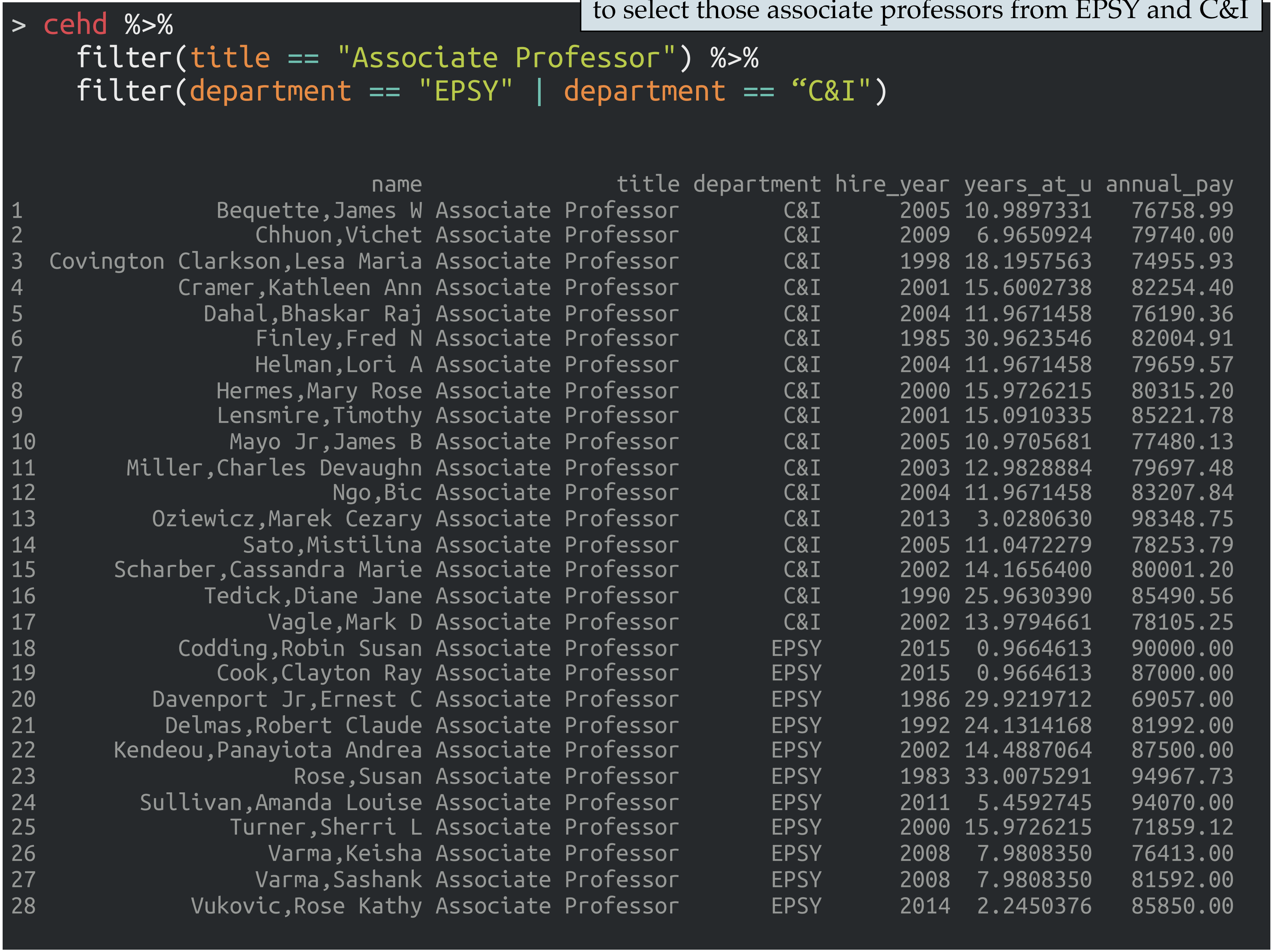
```
> cehd %>% filter(department == “EPSY" | title == "Associate Professor")


                            name               title department hire_year years_at_u annual_pay
1                Karatekin,Canan Associate Professor        ICD      1998 17.9274470   68732.93
2                 Koenig,Melissa Associate Professor        ICD      2007  8.9774127   88926.60
3               Bequette,James W Associate Professor        C&I      2005 10.9897331   76758.99
4                  Chhuon,Vichet Associate Professor        C&I      2009  6.9650924   79740.00
5  Covington Clarkson,Lesa Maria Associate Professor        C&I      1998 18.1957563   74955.93
6            Cramer,Kathleen Ann Associate Professor        C&I      2001 15.6002738   82254.40
                      …              …         …             …         …    …            …
90        Merighi,Joseph Richard Associate Professor        SSW      2013  2.9787817   82537.40
91         Reinardy,James Robert Associate Professor        SSW      1993 23.0061602  111499.25
92       Renner,Lynette Michelle Associate Professor        SSW      2013  2.9787817   84164.53
93         Shannon,Patricia Jean Associate Professor        SSW      2009  7.4633812   74227.44
94          VeLure Roholt,Ross R Associate Professor        SSW      1996 19.9972622   79018.94
```

Here to be selected the faculty member needs to be from the Educational Psychology department…OR..be an associate professor.

Your Turn

Select only the associate professors from C&I and Educational Psychology.

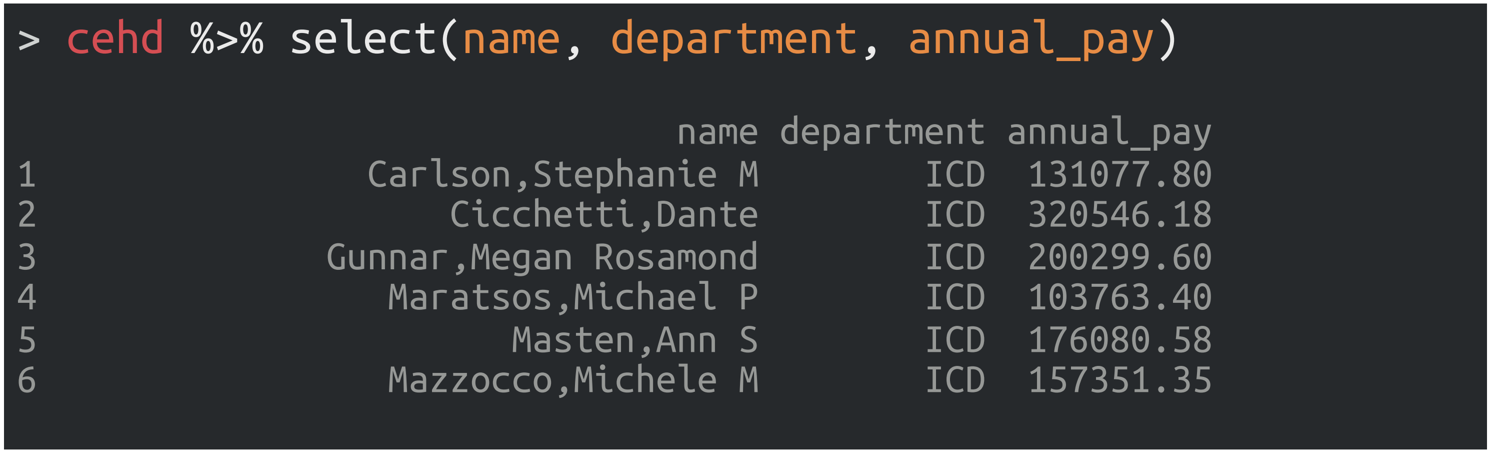
Here the key is to filter on title, then use another filter to select those associate professors from EPSY and C&I



```
> cehd %>%
    filter(title == "Associate Professor") %>%
    filter(department == "EPSY" | department == “C&I")


                            name               title department hire_year years_at_u annual_pay
1               Bequette,James W Associate Professor        C&I      2005 10.9897331   76758.99
2                  Chhuon,Vichet Associate Professor        C&I      2009  6.9650924   79740.00
3  Covington Clarkson,Lesa Maria Associate Professor        C&I      1998 18.1957563   74955.93
4            Cramer,Kathleen Ann Associate Professor        C&I      2001 15.6002738   82254.40
5              Dahal,Bhaskar Raj Associate Professor        C&I      2004 11.9671458   76190.36
6                  Finley,Fred N Associate Professor        C&I      1985 30.9623546   82004.91
7                  Helman,Lori A Associate Professor        C&I      2004 11.9671458   79659.57
8               Hermes,Mary Rose Associate Professor        C&I      2000 15.9726215   80315.20
9               Lensmire,Timothy Associate Professor        C&I      2001 15.0910335   85221.78
10               Mayo Jr,James B Associate Professor        C&I      2005 10.9705681   77480.13
11       Miller,Charles Devaughn Associate Professor        C&I      2003 12.9828884   79697.48
12                       Ngo,Bic Associate Professor        C&I      2004 11.9671458   83207.84
13         Oziewicz,Marek Cezary Associate Professor        C&I      2013  3.0280630   98348.75
14                Sato,Mistilina Associate Professor        C&I      2005 11.0472279   78253.79
15      Scharber,Cassandra Marie Associate Professor        C&I      2002 14.1656400   80001.20
16             Tedick,Diane Jane Associate Professor        C&I      1990 25.9630390   85490.56
17                  Vagle,Mark D Associate Professor        C&I      2002 13.9794661   78105.25
18           Codding,Robin Susan Associate Professor       EPSY      2015  0.9664613   90000.00
19              Cook,Clayton Ray Associate Professor       EPSY      2015  0.9664613   87000.00
20         Davenport Jr,Ernest C Associate Professor       EPSY      1986 29.9219712   69057.00
21          Delmas,Robert Claude Associate Professor       EPSY      1992 24.1314168   81992.00
22      Kendeou,Panayiota Andrea Associate Professor       EPSY      2002 14.4887064   87500.00
23                    Rose,Susan Associate Professor       EPSY      1983 33.0075291   94967.73
24        Sullivan,Amanda Louise Associate Professor       EPSY      2011  5.4592745   94070.00
25               Turner,Sherri L Associate Professor       EPSY      2000 15.9726215   71859.12
26                  Varma,Keisha Associate Professor       EPSY      2008  7.9808350   76413.00
27                 Varma,Sashank Associate Professor       EPSY      2008  7.9808350   81592.00
28            Vukovic,Rose Kathy Associate Professor       EPSY      2014  2.2450376   85850.00
```

Select a Subset of Columns

To select a subset of columns, we will use the `select()` function. The argument(s) for this function are column names of the data frame that you want to select.

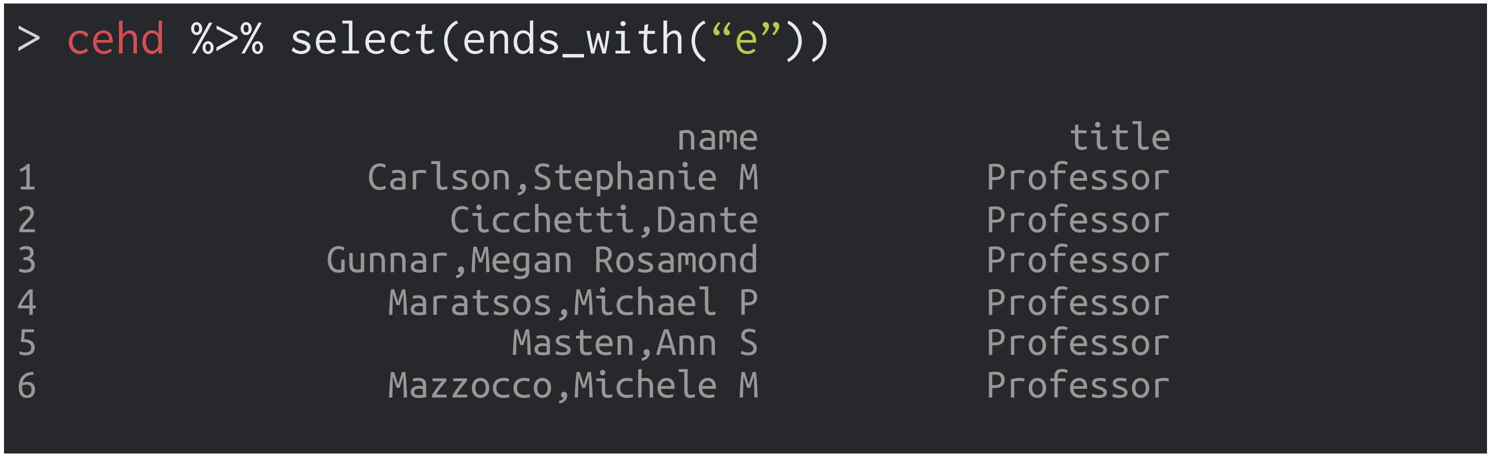


```
> cehd %>% select(name, department, annual_pay)

                                name department annual_pay
1                Carlson,Stephanie M        ICD  131077.80
2                    Cicchetti,Dante        ICD  320546.18
3              Gunnar,Megan Rosamond        ICD  200299.60
4                 Maratsos,Michael P        ICD  103763.40
5                       Masten,Ann S        ICD  176080.58
6                 Mazzocco,Michele M        ICD  157351.35
```

You can rename a column by naming it in the `select()` function. Here we rename *department* to *dept*.

```
> cehd %>% select(name, dept = department, annual_pay)

                                name dept annual_pay
1                Carlson,Stephanie M  ICD  131077.80
2                    Cicchetti,Dante  ICD  320546.18
3              Gunnar,Megan Rosamond  ICD  200299.60
4                 Maratsos,Michael P  ICD  103763.40
5                       Masten,Ann S  ICD  176080.58
6                 Mazzocco,Michele M  ICD  157351.35
```

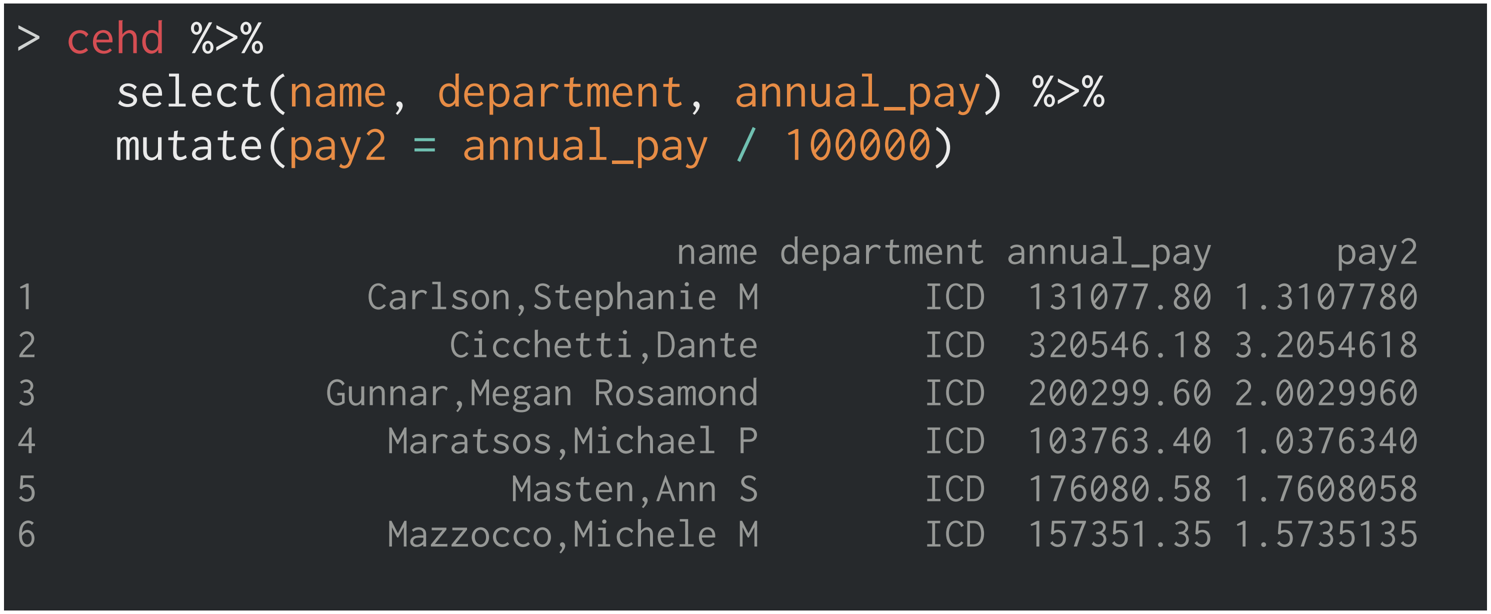
There are a number of helper functions you can use within `select()`. For example, `starts_with()`, `ends_with()`, and `contains()`. These let you quickly match larger blocks of variables that meet some criterion.



```
> cehd %>% select(ends_with(“e”))

                    name            title
1  Carlson,Stephanie M        Professor
2       Cicchetti,Dante        Professor
3  Gunnar,Megan Rosamond       Professor
4     Maratsos,Michael P       Professor
5           Masten,Ann S       Professor
6     Mazzocco,Michele M       Professor
```

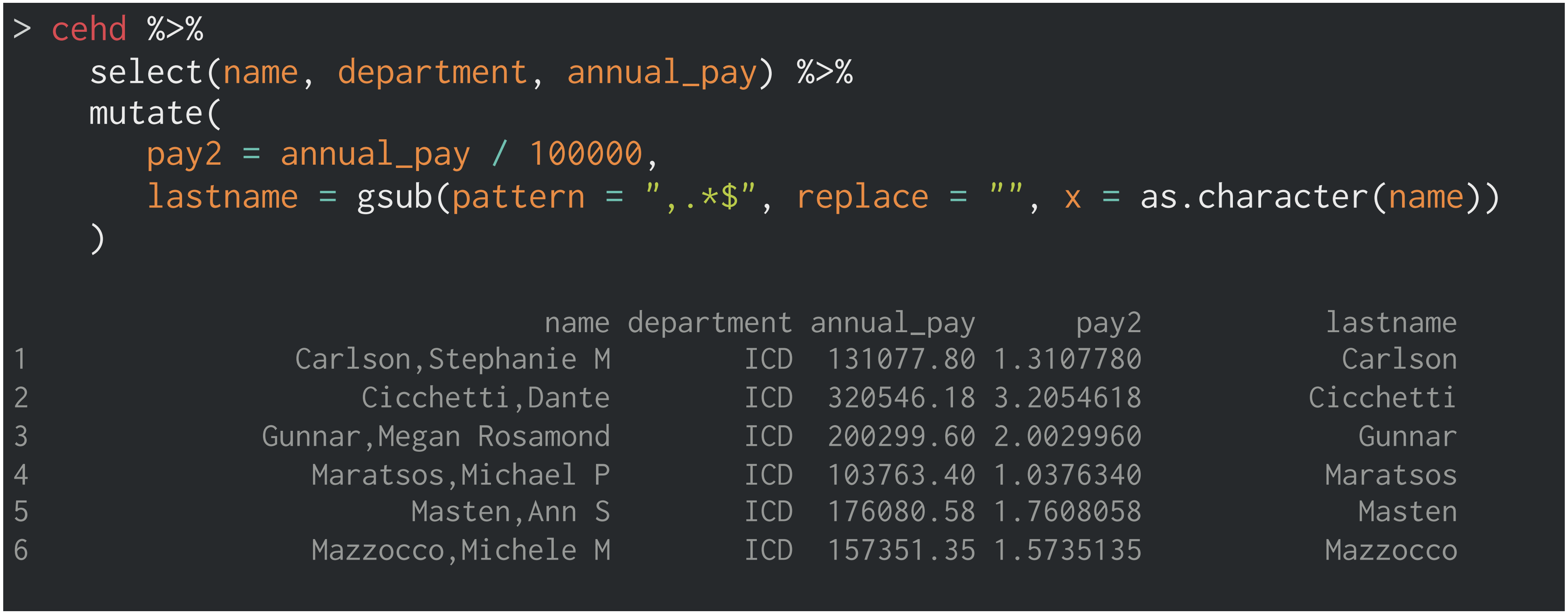
Create New Columns

To create new columns, we will use the `mutate()` function.



```
> cehd %>%
    select(name, department, annual_pay) %>%
    mutate(pay2 = annual_pay / 100000)

                                name department annual_pay      pay2
1                Carlson,Stephanie M        ICD  131077.80 1.3107780
2                    Cicchetti,Dante        ICD  320546.18 3.2054618
3              Gunnar,Megan Rosamond        ICD  200299.60 2.0029960
4                 Maratsos,Michael P        ICD  103763.40 1.0376340
5                       Masten,Ann S        ICD  176080.58 1.7608058
6                 Mazzocco,Michele M        ICD  157351.35 1.5735135
```

Create multiple columns by including each new column as an argument in the `mutate()` function.

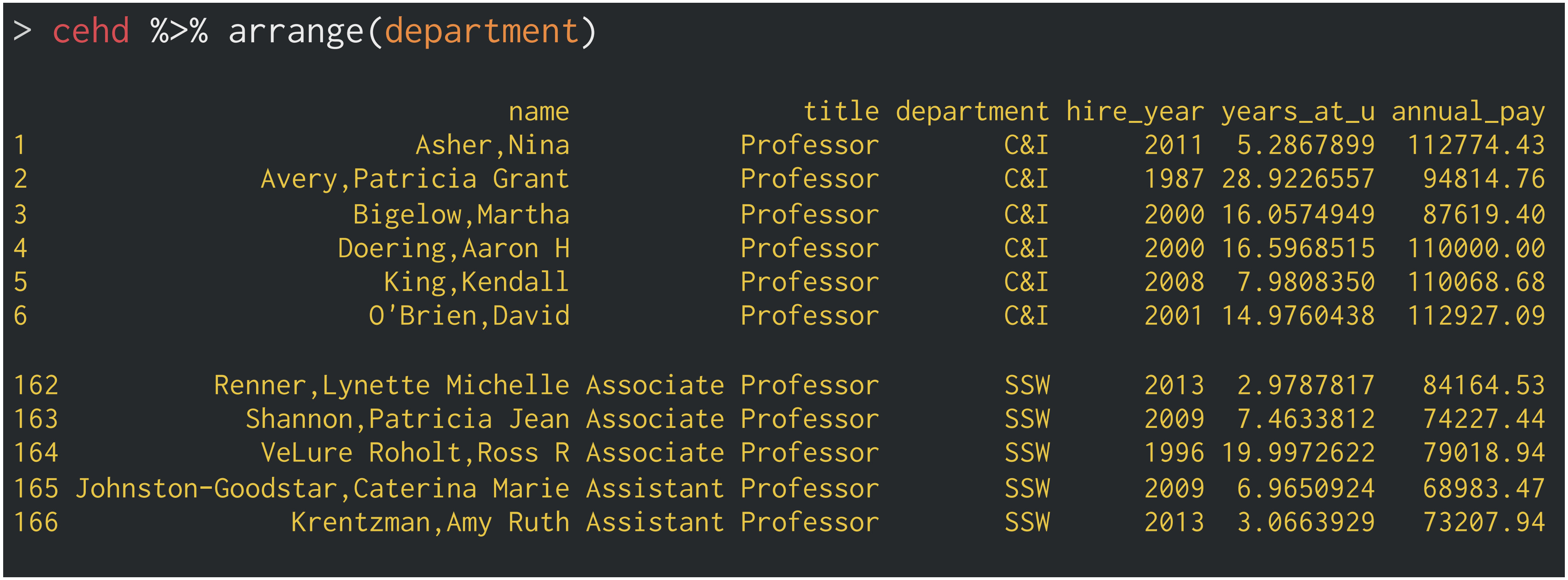


```
> cehd %>%
    select(name, department, annual_pay) %>%
    mutate(
       pay2 = annual_pay / 100000,
       lastname = gsub(pattern = ",.*$", replace = "", x = as.character(name))
    )

                        name department annual_pay     pay2      lastname
1         Carlson,Stephanie M        ICD  131077.80 1.3107780       Carlson
2             Cicchetti,Dante        ICD  320546.18 3.2054618     Cicchetti
3       Gunnar,Megan Rosamond        ICD  200299.60 2.0029960        Gunnar
4          Maratsos,Michael P        ICD  103763.40 1.0376340      Maratsos
5                Masten,Ann S        ICD  176080.58 1.7608058        Masten
6          Mazzocco,Michele M        ICD  157351.35 1.5735135      Mazzocco
```

Don’t worry about the syntax in the `gsub()` function. Here I am just illustrating how to create a two new variables using `mutate()`.

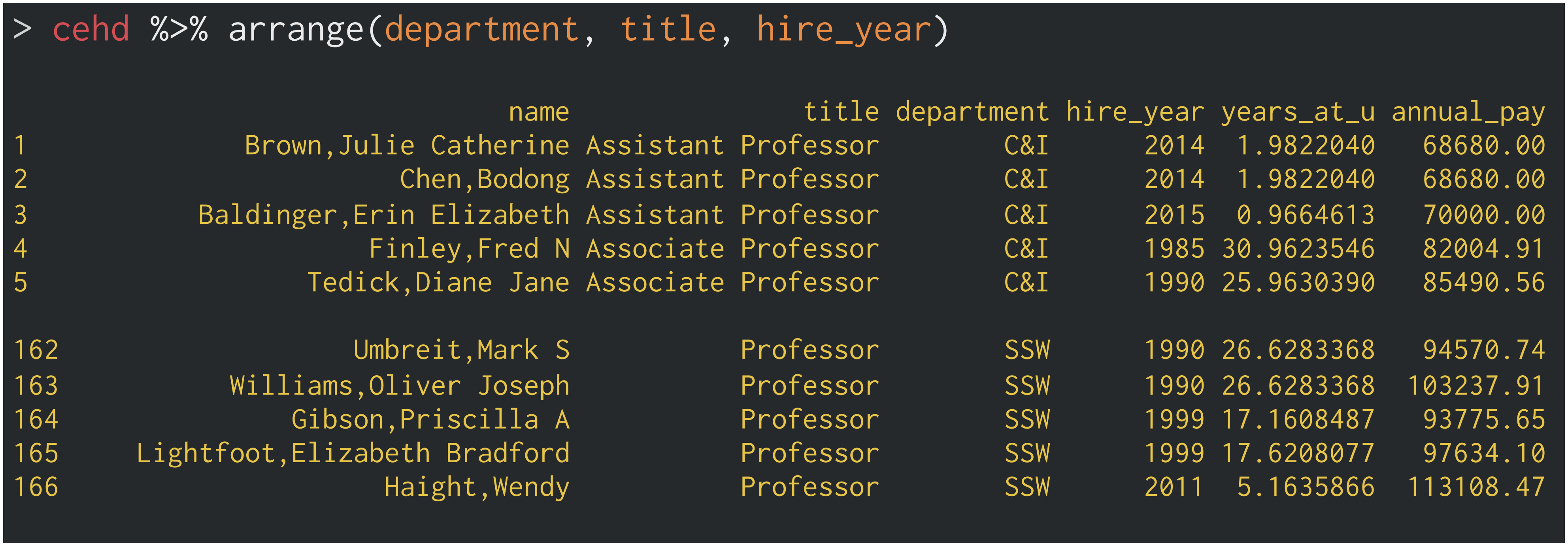
Arrange

The arrange() function sorts the data within a data frame. The data is ordered based on the column name provided in the argument(s).



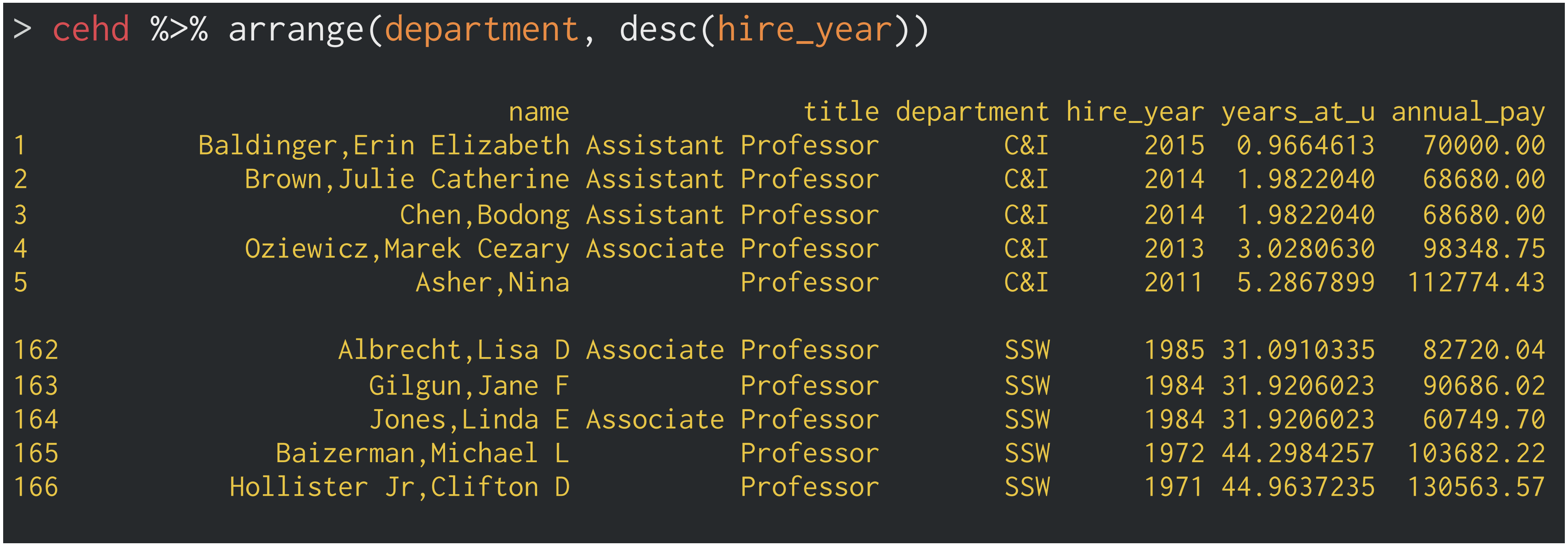
```
> cehd %>% arrange(department)

                                name               title department hire_year years_at_u annual_pay
1                         Asher,Nina           Professor        C&I      2011  5.2867899  112774.43
2               Avery,Patricia Grant           Professor        C&I      1987 28.9226557   94814.76
3                     Bigelow,Martha           Professor        C&I      2000 16.0574949   87619.40
4                    Doering,Aaron H           Professor        C&I      2000 16.5968515  110000.00
5                       King,Kendall           Professor        C&I      2008  7.9808350  110068.68
6                      O'Brien,David           Professor        C&I      2001 14.9760438  112927.09

162          Renner,Lynette Michelle Associate Professor        SSW      2013  2.9787817   84164.53
163            Shannon,Patricia Jean Associate Professor        SSW      2009  7.4633812   74227.44
164             VeLure Roholt,Ross R Associate Professor        SSW      1996 19.9972622   79018.94
165 Johnston-Goodstar,Caterina Marie Assistant Professor        SSW      2009  6.9650924   68983.47
166               Krentzman,Amy Ruth Assistant Professor        SSW      2013  3.0663929   73207.94
```

Multiple arguments sort first by the first argument, and then by the subsequent arguments.



```
> cehd %>% arrange(department, title, hire_year)

                                name               title department hire_year years_at_u annual_pay
1              Brown,Julie Catherine Assistant Professor        C&I      2014  1.9822040   68680.00
2                        Chen,Bodong Assistant Professor        C&I      2014  1.9822040   68680.00
3           Baldinger,Erin Elizabeth Assistant Professor        C&I      2015  0.9664613   70000.00
4                      Finley,Fred N Associate Professor        C&I      1985 30.9623546   82004.91
5                  Tedick,Diane Jane Associate Professor        C&I      1990 25.9630390   85490.56

162                   Umbreit,Mark S           Professor        SSW      1990 26.6283368   94570.74
163           Williams,Oliver Joseph           Professor        SSW      1990 26.6283368  103237.91
164               Gibson,Priscilla A           Professor        SSW      1999 17.1608487   93775.65
165     Lightfoot,Elizabeth Bradford           Professor        SSW      1999 17.6208077   97634.10
166                     Haight,Wendy           Professor        SSW      2011  5.1635866  113108.47
```

Use the `desc()` function to order a column in descending order:



```
> cehd %>% arrange(department, desc(hire_year))

                                name               title department hire_year years_at_u annual_pay
1           Baldinger,Erin Elizabeth Assistant Professor        C&I      2015  0.9664613   70000.00
2              Brown,Julie Catherine Assistant Professor        C&I      2014  1.9822040   68680.00
3                        Chen,Bodong Assistant Professor        C&I      2014  1.9822040   68680.00
4              Oziewicz,Marek Cezary Associate Professor        C&I      2013  3.0280630   98348.75
5                         Asher,Nina           Professor        C&I      2011  5.2867899  112774.43

162                  Albrecht,Lisa D Associate Professor        SSW      1985 31.0910335   82720.04
163                    Gilgun,Jane F           Professor        SSW      1984 31.9206023   90686.02
164                    Jones,Linda E Associate Professor        SSW      1984 31.9206023   60749.70
165              Baizerman,Michael L           Professor        SSW      1972 44.2984257  103682.22
166           Hollister Jr,Clifton D           Professor        SSW      1971 44.9637235  130563.57
```

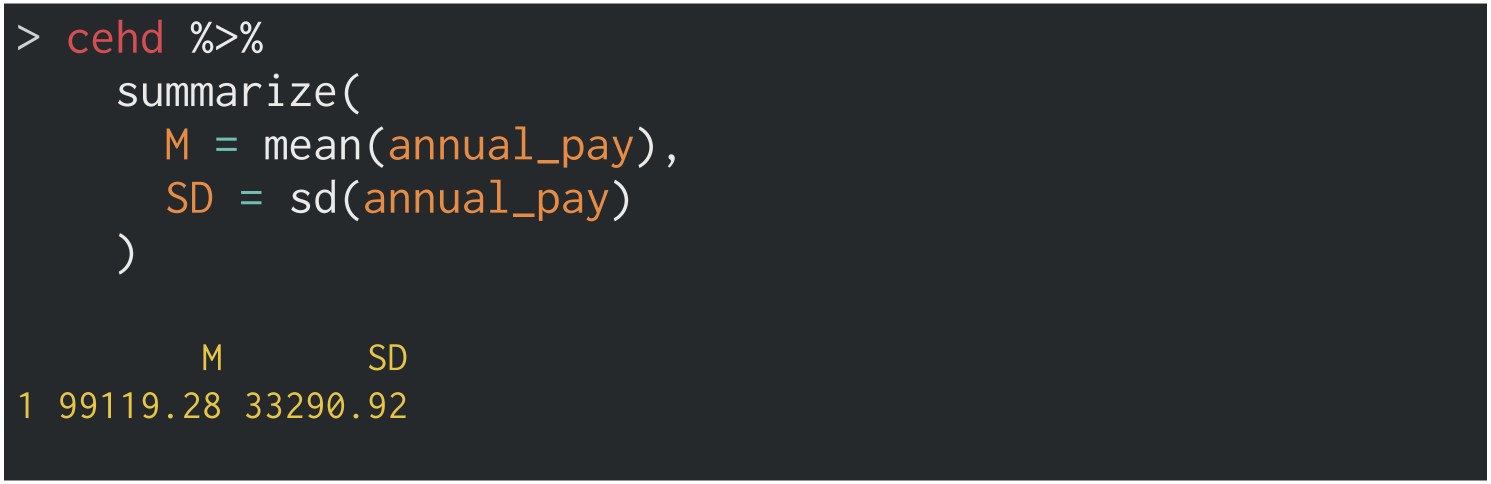
Summarizing

The `summarize()` function is used to compute summaries of data. It collapses a data frame to a single row.

```
> cehd %>% summarize(M = mean(annual_pay))

         M
1 99119.28
```

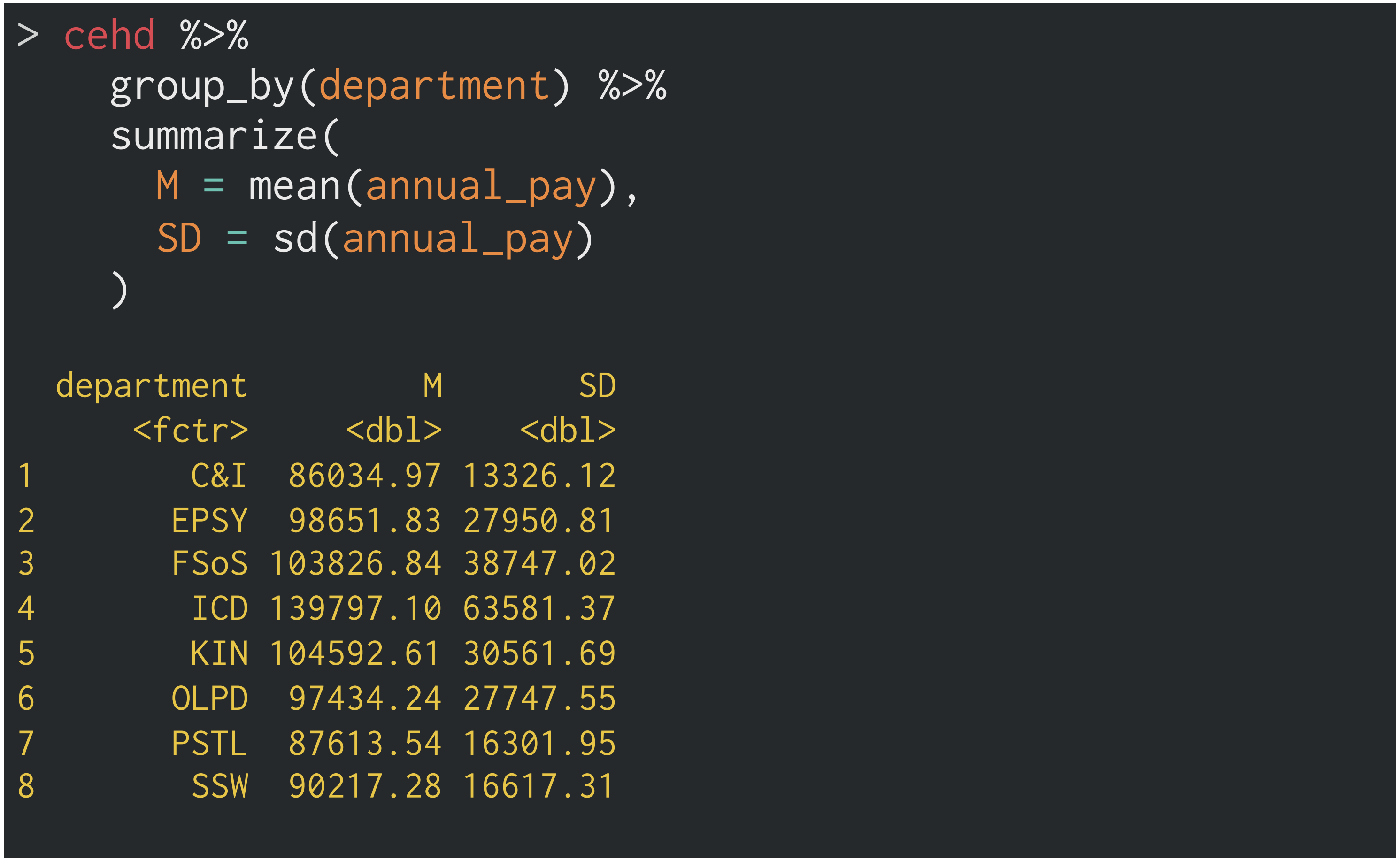
The output is a single row data frame with a column called *M*.

Multiple summaries can be computed by providing more than one argument to the `summarize()` function. The output is still a single row data frame, but now there will be multiple columns, one for each summary computation.



```
> cehd %>%
    summarize(
      M = mean(annual_pay),
      SD = sd(annual_pay)
    )

         M       SD
1 99119.28 33290.92
```

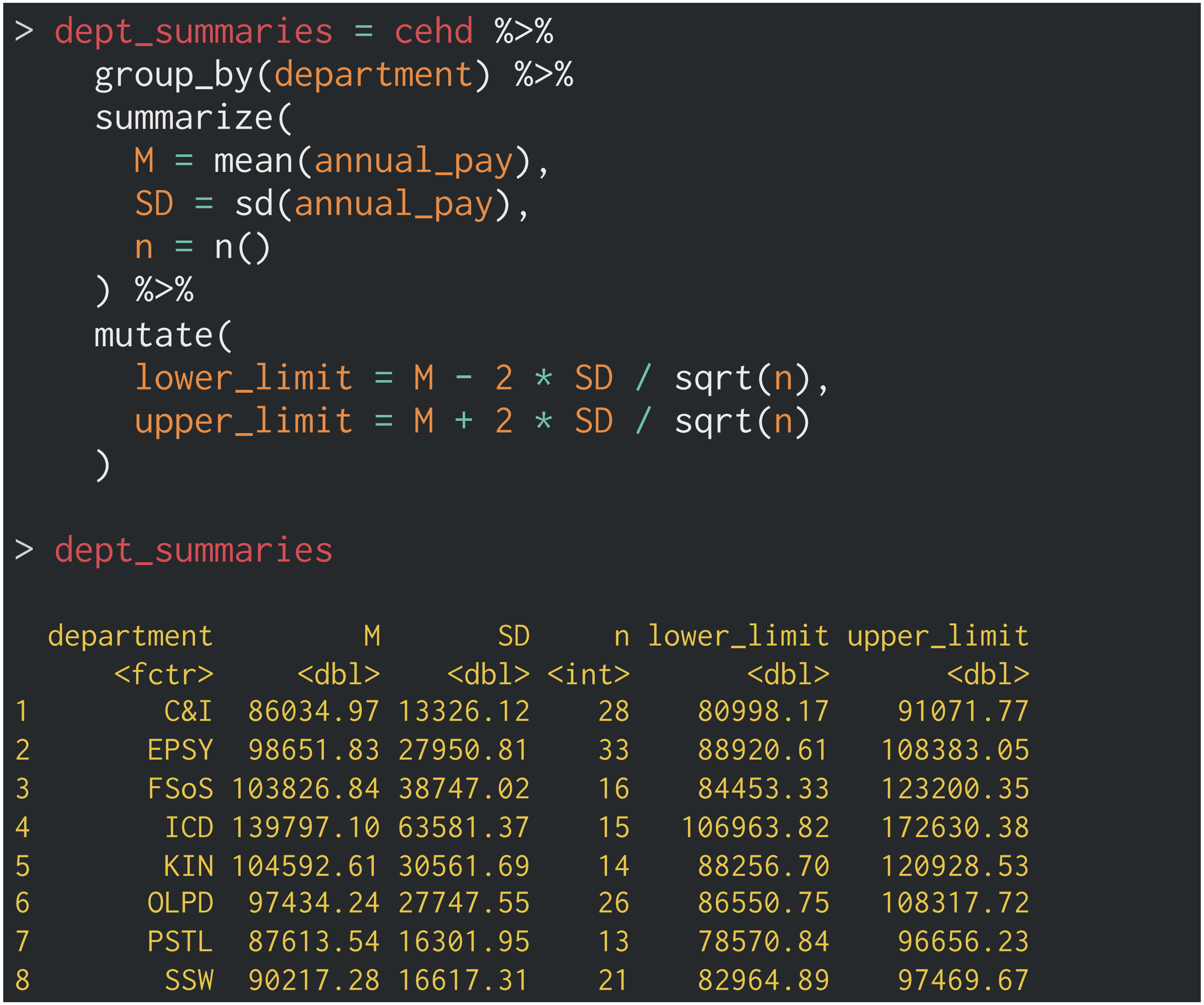
The `group_by()` function groups the data by a specified variable. By itself, it does nothing, but it is powerful when the grouped output is chained to other functions, such as `summarize()`.



```
> cehd %>%
    group_by(department) %>%
    summarize(
      M = mean(annual_pay),
      SD = sd(annual_pay)
    )

  department         M       SD
      <fctr>     <dbl>    <dbl>
1        C&I  86034.97 13326.12
2       EPSY  98651.83 27950.81
3       FSoS 103826.84 38747.02
4        ICD 139797.10 63581.37
5        KIN 104592.61 30561.69
6       OLPD  97434.24 27747.55
7       PSTL  87613.54 16301.95
8        SSW  90217.28 16617.31
```

The output from summarize now has multiple rows, one for each level of the grouping variable. It computes the two summaries for each of those groups and outputs this in a data frame. It also includes a column for the grouping level.

Here we group by department, compute the mean salary, the standard deviation of the salaries and the departmental count. Then we use the mutate() function to compute the upper and lower limits for the 95% CI (using the formula for the SEM).



```
> dept_summaries = cehd %>%
    group_by(department) %>%
    summarize(
      M = mean(annual_pay),
      SD = sd(annual_pay),
      n = n()
    ) %>%
    mutate(
      lower_limit = M - 2 * SD / sqrt(n),
      upper_limit = M + 2 * SD / sqrt(n)
    )

> dept_summaries

  department         M       SD     n lower_limit upper_limit
      <fctr>     <dbl>    <dbl> <int>       <dbl>       <dbl>
1        C&I  86034.97 13326.12    28    80998.17    91071.77
2       EPSY  98651.83 27950.81    33    88920.61   108383.05
3       FSoS 103826.84 38747.02    16    84453.33   123200.35
4        ICD 139797.10 63581.37    15   106963.82   172630.38
5        KIN 104592.61 30561.69    14    88256.70   120928.53
6       OLPD  97434.24 27747.55    26    86550.75   108317.72
7       PSTL  87613.54 16301.95    13    78570.84    96656.23
8        SSW  90217.28 16617.31    21    82964.89    97469.67
```

Since the output of any dplyr call is a data frame, we can use this in a ggplot. Typically one would save the output from dplyr into an object and then call that in the ggplot syntax.

```
> library(ggplot2)

> ggplot(data = dept_summaries, aes(x = department, y = M)) +
    geom_segment(aes(x = department, xend = department,
                     y = lower_limit, yend = upper_limit)) +
    geom_line(aes(group = 1), linetype = "dotted") +
    geom_point() +
    theme_bw() +
    xlab("") +
    ylab("Average Base Salary")
```

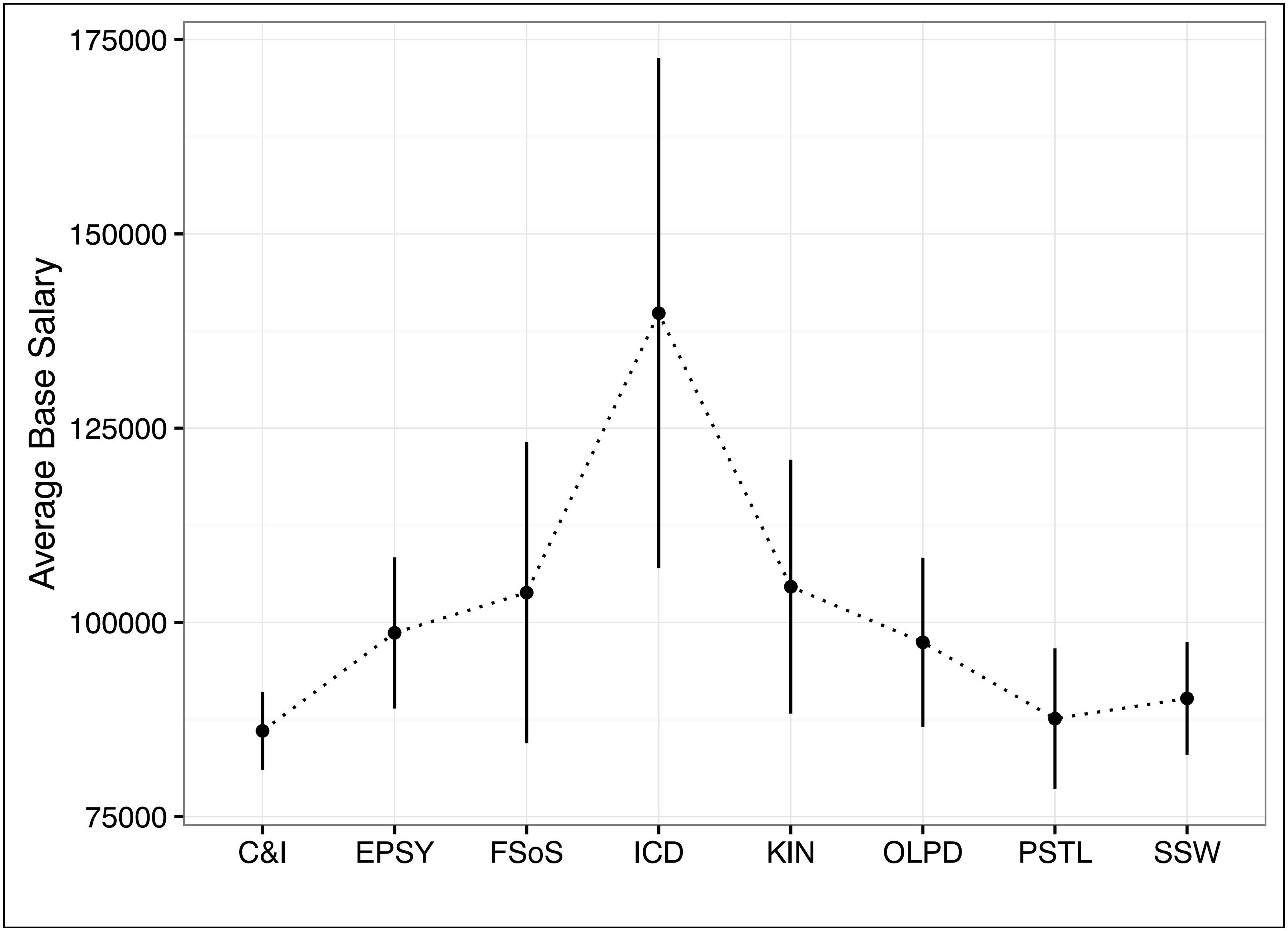


Figure 1. Average base salary for faculty in CEHD by department. The 95% CI for the mean is also included.

dplyr Resources

- **dplyr Cheatsheet**: A one-page (front and back) cheatsheet of dplyr syntax with pictures https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

- **Introduction to dplyr:** Web-based tutorial with examples https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

- **tidy data paper:** A paper that outlines how to tidy/clean data for analysis. http://vita.had.co.nz/papers/tidy-data.html

`#protip`: Use Google to find out how to do just about anything with dplyr.