Project Draft 4 Pie Plot

Carrie Hashimoto

```
## Import data
library(ggplot2)
library(stringr)
library(RSQLite)

dcon <- dbConnect(SQLite(), dbname = "projectDB.db")
dbListTables(dcon)
```

```
## [1] "tableCrashes"
```

```
dbListFields(dcon, "tableCrashes")
```

```
##  [1] "CRASHDATE"                  "CRASHTIME"
##  [3] "BOROUGH"                    "ZIPCODE"
##  [5] "LATITUDE"                   "LONGITUDE"
##  [7] "LOCATION"                   "ONSTREETNAME"
##  [9] "CROSSSTREETNAME"            "OFFSTREETNAME"
## [11] "NUMBEROFPERSONSINJURED"     "NUMBEROFPERSONSKILLED"
## [13] "NUMBEROFPEDESTRIANSINJURED" "NUMBEROFPEDESTRIANSKILLED"
## [15] "NUMBEROFCYCLISTINJURED"     "NUMBEROFCYCLISTKILLED"
## [17] "NUMBEROFMOTORISTINJURED"    "NUMBEROFMOTORISTKILLED"
## [19] "CONTRIBUTINGFACTORVEHICLE1" "CONTRIBUTINGFACTORVEHICLE2"
## [21] "CONTRIBUTINGFACTORVEHICLE3" "CONTRIBUTINGFACTORVEHICLE4"
## [23] "CONTRIBUTINGFACTORVEHICLE5" "COLLISION_ID"
## [25] "VEHICLETYPECODE1"           "VEHICLETYPECODE2"
## [27] "VEHICLETYPECODE3"           "VEHICLETYPECODE4"
## [29] "VEHICLETYPECODE5"
```

```
# ## Code to get cleaned table, for reference
# res <- dbSendQuery(conn = dcon, "
# SELECT COLLISION_ID,
#   substr(CRASHDATE,7,10) as year,
#   CASE
#     WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Cell Phone (hands-free)',
#       'Cell Phone (hand-Held)',
#       'Other Electronic Device','Texting','Cell Phone (hand-held)')
#       THEN 'Cell phone/electronic'
#     WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Illnes','Illness',
#       'Physical Disability','Prescription medication')
#       THEN 'Medical'
#     WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Drugs (illegal)' THEN 'Drugs (Illegal)'
#     WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Reaction to Other Uninvolved Vehicle'
#       THEN 'Reaction to Uninvolved Vehicle'
```

```
#     WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Fatigued/Drowsy','Fell Asleep',
#      'Lost Consciousness')
#       THEN 'Fatigued/drowsy/asleep/unconscious'
#     WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('80','1')
#       THEN 'Unknown'
#     WHEN CONTRIBUTINGFACTORVEHICLE1 IS NULL -- (to be consistent w last time)
#       THEN 'Unknown'
#     ELSE CONTRIBUTINGFACTORVEHICLE1
#     END AS MODCONTRIBUTINGFACTORVEHICLE1
# FROM tableCrashes;
# ")
# tableCrashesCleaned <- dbFetch(res, -1)
# dbClearResult(res)

## Code to get ten contributing factors that result in the most injuries
# res <- dbSendQuery(conn = dcon, "
# SELECT MODCONTRIBUTINGFACTORVEHICLE1, SUM(NUMBEROFPERSONSINJURED) as COUNT_INJURED FROM
#   (SELECT COLLISION_ID,
#     substr(CRASHDATE,7,10) as year, NUMBEROFPERSONSINJURED,
#     CASE
#       WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Cell Phone (hands-free)',
#         'Cell Phone (hand-Held)',
#         'Other Electronic Device','Texting','Cell Phone (hand-held)')
#         THEN 'Cell phone/electronic'
#       WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Illnes','Illness',
#         'Physical Disability','Prescription medication')
#         THEN 'Medical'
#       WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Drugs (illegal)' THEN 'Drugs (Illegal)'
#       WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Reaction to Other Uninvolved Vehicle'
#         THEN 'Reaction to Uninvolved Vehicle'
#       WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Fatigued/Drowsy',
#         'Fell Asleep','Lost Consciousness')
#         THEN 'Fatigued/drowsy/asleep/unconscious'
#       WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('80','1','Unspecified','Other Vehicular')
#         THEN 'Unknown'
#       WHEN CONTRIBUTINGFACTORVEHICLE1 IS NULL -- (to be consistent w last time)
#         THEN 'Unknown'
#       ELSE CONTRIBUTINGFACTORVEHICLE1
#       END AS MODCONTRIBUTINGFACTORVEHICLE1
#   FROM tableCrashes)
# WHERE MODCONTRIBUTINGFACTORVEHICLE1 NOT IN
#   (SELECT DISTINCT CONTRIBUTINGFACTORVEHICLE1
#   FROM tableCrashes
#   WHERE NUMBEROFPERSONSINJURED IS NULL
#   AND CONTRIBUTINGFACTORVEHICLE1 NOT NULL)
#   AND MODCONTRIBUTINGFACTORVEHICLE1 != 'Unknown'
# GROUP BY MODCONTRIBUTINGFACTORVEHICLE1
# ORDER BY COUNT_INJURED DESC
# LIMIT 10;
# ")
# ten_contr_factors_counts <- dbFetch(res, -1)
# dbClearResult(res)
```

```
## Crashes subsetted by ten most injurious contributing factors
# res <- dbSendQuery(conn = dcon, "
# SELECT CAST(substr(CRASHDATE,7,10) AS INT) AS year, MODCONTRIBUTINGFACTORVEHICLE1
# FROM (
#     SELECT CRASHDATE,
#         CASE
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Cell Phone (hands-free)',
#             'Cell Phone (hand-Held)',
#             'Other Electronic Device','Texting','Cell Phone (hand-held)')
#             THEN 'Cell phone/electronic'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Illnes','Illness',
#             'Physical Disability','Prescription medication')
#             THEN 'Medical'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Drugs (illegal)'
#             THEN 'Drugs (Illegal)'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Reaction to Other Uninvolved Vehicle'
#             THEN 'Reaction to Uninvolved Vehicle'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Fatigued/Drowsy',
#             'Fell Asleep','Lost Consciousness')
#             THEN 'Fatigued/drowsy/asleep/unconscious'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('80','1','Unspecified',
#             'Other Vehicular')
#             THEN 'Unknown'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IS NULL
#             -- (to be consistent w last time)
#             THEN 'Unknown'
#           ELSE CONTRIBUTINGFACTORVEHICLE1
#           END AS MODCONTRIBUTINGFACTORVEHICLE1
#     FROM tableCrashes
#     WHERE NUMBEROFPERSONSINJURED NOT NULL)
# WHERE MODCONTRIBUTINGFACTORVEHICLE1 IN (
#   SELECT MODCONTRIBUTINGFACTORVEHICLE1 FROM (
#     SELECT MODCONTRIBUTINGFACTORVEHICLE1, SUM(NUMBEROFPERSONSINJURED)
#     AS COUNT_INJURED FROM
#       (SELECT COLLISION_ID,
#         CAST(substr(CRASHDATE,7,10) AS INT) AS year, NUMBEROFPERSONSINJURED,
#         CASE
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Cell Phone (hands-free)',
#             'Cell Phone (hand-Held)',
#             'Other Electronic Device','Texting','Cell Phone (hand-held)')
#             THEN 'Cell phone/electronic'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Illnes','Illness',
#             'Physical Disability','Prescription medication')
#             THEN 'Medical'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Drugs (illegal)'
#             THEN 'Drugs (Illegal)'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Reaction to Other Uninvolved Vehicle'
#             THEN 'Reaction to Uninvolved Vehicle'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Fatigued/Drowsy',
#             'Fell Asleep','Lost Consciousness')
#             THEN 'Fatigued/drowsy/asleep/unconscious'
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('80','1','Unspecified','Other Vehicular')
#             THEN 'Unknown'
```

```
#           WHEN CONTRIBUTINGFACTORVEHICLE1 IS NULL
#             THEN 'Unknown'
#           ELSE CONTRIBUTINGFACTORVEHICLE1
#           END AS MODCONTRIBUTINGFACTORVEHICLE1
#       FROM tableCrashes
#       )
#     WHERE MODCONTRIBUTINGFACTORVEHICLE1 NOT IN
#       (SELECT DISTINCT CONTRIBUTINGFACTORVEHICLE1
#       FROM tableCrashes
#       WHERE NUMBEROFPERSONSINJURED IS NULL
#       AND CONTRIBUTINGFACTORVEHICLE1 NOT NULL)
#     AND MODCONTRIBUTINGFACTORVEHICLE1 != 'Unknown'
#     GROUP BY MODCONTRIBUTINGFACTORVEHICLE1
#     ORDER BY COUNT_INJURED DESC
#     LIMIT 10)
#   );
# ")
# tableCrashes_ten_contr_factors <- dbFetch(res, -1)
# dbClearResult(res)


res <- dbSendQuery(conn = dcon, "
CREATE VIEW tableCrashesSubset AS SELECT CAST(substr(CRASHDATE,7,10) AS INT)
AS year, MODCONTRIBUTINGFACTORVEHICLE1, COUNT(*) AS COUNT
FROM (
    SELECT CRASHDATE,
        CASE
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Cell Phone (hands-free)',
            'Cell Phone (hand-Held)',
            'Other Electronic Device','Texting','Cell Phone (hand-held)')
            THEN 'Cell phone/electronic'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Illnes','Illness',
            'Physical Disability','Prescription medication')
            THEN 'Medical'
          WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Drugs (illegal)'
            THEN 'Drugs (Illegal)'
          WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Reaction to Other Uninvolved Vehicle'
            THEN 'Reaction to Uninvolved Vehicle'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Fatigued/Drowsy',
            'Fell Asleep','Lost Consciousness')
            THEN 'Fatigued/drowsy/asleep/unconscious'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('80','1','Unspecified',
            'Other Vehicular')
            THEN 'Unknown'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IS NULL
            -- (to be consistent w last time)
            THEN 'Unknown'
          ELSE CONTRIBUTINGFACTORVEHICLE1
          END AS MODCONTRIBUTINGFACTORVEHICLE1
    FROM tableCrashes
    WHERE NUMBEROFPERSONSINJURED NOT NULL)
WHERE MODCONTRIBUTINGFACTORVEHICLE1 IN (
  SELECT MODCONTRIBUTINGFACTORVEHICLE1 FROM (
```

```
    SELECT MODCONTRIBUTINGFACTORVEHICLE1, SUM(NUMBEROFPERSONSINJURED)
    AS COUNT_INJURED FROM
      (SELECT COLLISION_ID,
        CAST(substr(CRASHDATE,7,10) AS INT) AS year, NUMBEROFPERSONSINJURED,
        CASE
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Cell Phone (hands-free)',
            'Cell Phone (hand-Held)',
            'Other Electronic Device','Texting','Cell Phone (hand-held)')
            THEN 'Cell phone/electronic'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Illnes','Illness',
            'Physical Disability','Prescription medication')
            THEN 'Medical'
          WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Drugs (illegal)'
            THEN 'Drugs (Illegal)'
          WHEN CONTRIBUTINGFACTORVEHICLE1 = 'Reaction to Other Uninvolved Vehicle'
            THEN 'Reaction to Uninvolved Vehicle'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('Fatigued/Drowsy',
            'Fell Asleep','Lost Consciousness')
            THEN 'Fatigued/drowsy/asleep/unconscious'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IN ('80','1','Unspecified','Other Vehicular')
            THEN 'Unknown'
          WHEN CONTRIBUTINGFACTORVEHICLE1 IS NULL
            THEN 'Unknown'
          ELSE CONTRIBUTINGFACTORVEHICLE1
          END AS MODCONTRIBUTINGFACTORVEHICLE1
        FROM tableCrashes
        )
    WHERE MODCONTRIBUTINGFACTORVEHICLE1 NOT IN
      (SELECT DISTINCT CONTRIBUTINGFACTORVEHICLE1
      FROM tableCrashes
      WHERE NUMBEROFPERSONSINJURED IS NULL
      AND CONTRIBUTINGFACTORVEHICLE1 NOT NULL)
    AND MODCONTRIBUTINGFACTORVEHICLE1 != 'Unknown'
    GROUP BY MODCONTRIBUTINGFACTORVEHICLE1
    ORDER BY COUNT_INJURED DESC
    LIMIT 10)
  )
GROUP BY year, MODCONTRIBUTINGFACTORVEHICLE1;
")

res <- dbSendQuery(conn = dcon, "
SELECT year, MODCONTRIBUTINGFACTORVEHICLE1 AS contributing_factor,
CAST(COUNT AS DOUBLE)/CAST(TOTAL AS DOUBLE)*100 AS percent
FROM tableCrashesSubset JOIN (
  SELECT year AS year_total, SUM(COUNT) as TOTAL
  FROM tableCrashesSubset
  GROUP BY year_total)
ON year = year_total
")
```

```
## Warning: Closing open result set, pending rows
```

```
tableCrashesPercentContributingByYear <- dbFetch(res, -1)
dbClearResult(res)


## Old methods
# data.all <- read.csv('Motor_Vehicle_Collisions_-_Crashes.csv')
# ## Feature engineering
# data.all[data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in%
#         c("Cell Phone (hands-free)","Cell Phone (hand-Held)",
#         "Other Electronic Device","Texting","Cell Phone (hand-held)"),
#         ]$CONTRIBUTING.FACTOR.VEHICLE.1 <- "Cell phone/electronic"
# data.all[data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in% c("Illnes","Illness",
#         "Physical Disability","Prescription medication"),
#         ]$CONTRIBUTING.FACTOR.VEHICLE.1 <- "Medical"
# data.all[data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in%
#         c("Drugs (illegal)"),]$CONTRIBUTING.FACTOR.VEHICLE.1 <- "Drugs (Illegal)"
# data.all[data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in%
#         c("Reaction to Other Uninvolved Vehicle"),
#         ]$CONTRIBUTING.FACTOR.VEHICLE.1 <- "Reaction to Uninvolved Vehicle"
# data.all[data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in%
#         c("Fatigued/Drowsy","Fell Asleep","Lost Consciousness"),
#         ]$CONTRIBUTING.FACTOR.VEHICLE.1 <- "Fatigued/drowsy/asleep/unconscious"
# data.all[data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in% c("80",
#         "1","Unspecified",""),]$CONTRIBUTING.FACTOR.VEHICLE.1 <- "Unknown"
#
# ## Get the number of crashes with each contributing factor
# contr_factor_counts <- aggregate(data.all$NUMBER.OF.PERSONS.INJURED,
#                         by=list(data.all$CONTRIBUTING.FACTOR.VEHICLE.1), FUN=sum)
#
# ## Find the 10 contributing factors with the most injuries
# top_10_contr_factors <- contr_factor_counts[order(-contr_factor_counts$x),][1:10,]
#
# ## Subset data by these factors
# car_crashes_top_factors <- subset(data.all,
#                             data.all$CONTRIBUTING.FACTOR.VEHICLE.1 %in%
#                             top_10_contr_factors$Group.1)
#
# ## Remove NAs in injured count column
# car_crashes_top_factors <-
# car_crashes_top_factors[!is.na(
#   car_crashes_top_factors$NUMBER.OF.PERSONS.INJURED),]
#
# ## Create year column so we can group by to check
# car_crashes_top_factors$year <- as.numeric(format(as.Date(
#   car_crashes_top_factors$CRASH.DATE,format='%m/%d/%Y'),"%Y"))


########## Code comparing tables to check that they're consistent
# library(dplyr)
#
# dim(tableCrashes_ten_contr_factors)
# dim(car_crashes_top_factors)
#
```

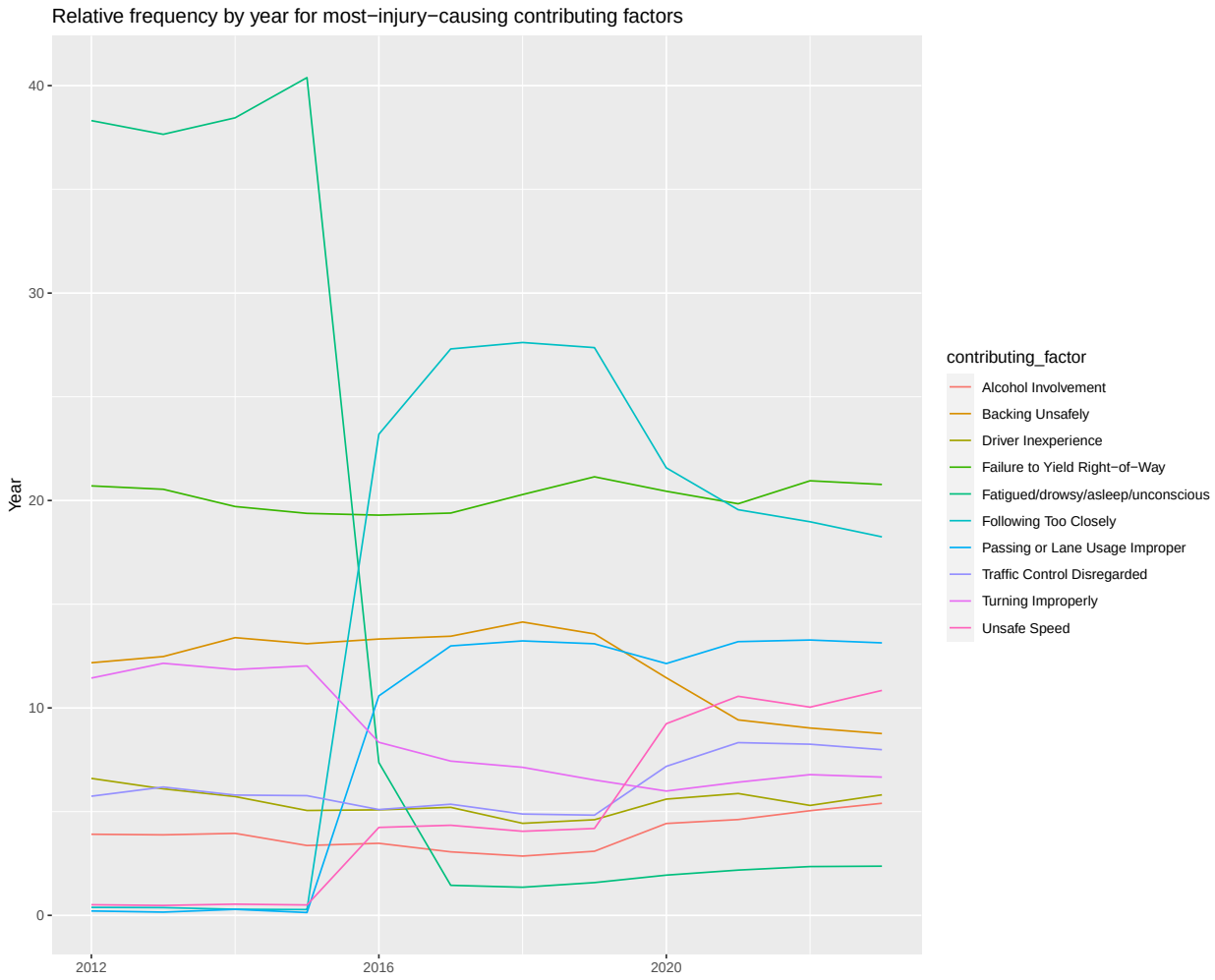
```
# ten_contr_factors_counts %>%
#   full_join(top_10_contr_factors,by=join_by(MODCONTRIBUTINGFACTORVEHICLE1==Group.1))
# # Same results
#
# car_crashes_top_factors %>%
#   group_by(year) %>%
#   summarise(count=n())
#
# tableCrashes_ten_contr_factors %>%
#   group_by(year) %>%
#   summarise(count=n())
```

```
### Pie chart plots for original version and SQL version
#png(file = "base_r_pieplot.png", width = 2000, height = 800, pointsize = 10)
# pie <- ggplot(car_crashes_top_factors, aes(x = factor(1),
#        fill = CONTRIBUTING.FACTOR.VEHICLE.1)) +
#        geom_bar(width = 1, position = "fill") +
#        coord_polar(theta = "y") + facet_wrap(.~ year,nrow=3) +
#        labs(title =
#       "Frequency by year for most-injury-causing contributing factors",
#        x = "",
#        y = "Year") +
#       scale_fill_discrete("Contributing factor")
# pie
#dev.off()
```

```
# png(file = "SQL_line_plot.png", width = 2000, height = 800, pointsize = 10)
# sql_line_plot <- ggplot(tableCrashes_ten_contr_factors, aes(x = factor(1),
#            fill = MODCONTRIBUTINGFACTORVEHICLE1)) +
#            geom_line() +
#            facet_wrap(.~ year,nrow=3) +
#            labs(title = "Frequency by year for most-injury-causing contributing factors",
#            x = "", y = "Year") +
#            scale_fill_discrete("Contributing factor")

library(scales)

ggplot(tableCrashesPercentContributingByYear) +
  aes(x=year, y=percent,color=contributing_factor) +
  geom_line() +
  labs(title = "Relative frequency by year for most-injury-causing contributing factors",
  x = "", y = "Year") +
  scale_fill_discrete("Contributing factor")
```




```
# dev.off()

dbSendQuery(conn = dcon, "
DROP VIEW tableCrashesSubset
")
```

```
## <SQLiteResult>
##   SQL
## DROP VIEW tableCrashesSubset
##
##   ROWS Fetched: 0 [complete]
##        Changed: 0
```

```
dbDisconnect(dcon)
```

```
## Warning in connection_release(conn@ptr): There are 1 result in use. The
## connection will be released when they are closed
```