פרויקט במדעי הנתונים - יעילות לשכות התעסוקה

אדריאנה שפטל, מידן גרינברג, אביחי קלנגל, ידיד זולדן וטל אלבז

אנו סטודנטים להנדסת תעשיה וניהול במכללת סמי שמעון (שנה ג') במגמת מערכות מידע. במסגרת קורס "מדעי הנתונים", וכן כחלק מרצוננו לפיתוח ראיה רחבה לגבי התעסוקה במדינת ישראל, בחרנו להשתמש ולנתח את בסיס הנתונים הממשלתי אשר מרכז את נתוני התעסוקה פר יישוב לאורך השנים 2010-2019. במסגרת הפרויקט ביצענו ניתוחים סטטיסטיים אשר מבצעים מיקוד על הנתונים / ההשערות שרצינו לבדוק עבור שאלת המחקר. כמו כן, בנינו גרפים אשר מציגים את בסיס הנתונים בצורה ויזואלית ונגישה לקורא.

מקור ההשראה שלנו היה התבוננות על העשור האחרון, במהלכו נצפתה בישראל ירידה באבטלה, עלייה בהשתתפות בשוק העבודה וגם עלייה בשכר – נתונים מעודדים בהחלט. נתונים אלה באים לצד התפתחות טכנולוגית ברשתות האינטרנט שהפכו את חיפוש העבודה לקל מתמיד. בפרויקט ננסה להראות את הקשר והתרומה של לשכות התעסוקה למצב שוק העבודה בישראל, דרך ניתוח כמות דורשי העבודה וכמות מקבלי העבודה דרך לשכות התעסוקה – בניתוחים לפי אזורים, ערים ושנים.

בסיס הנתונים

בחרנו להשתמש בבסיס נתונים קיים, מתוך אתר מאגרי המידע הממשלתיים של ישראל (https://data.gov.il/dataset)

בסיס הנתונים שבחרנו מכיל מידע ברמת ישוב לפי שנים (https://data.gov.il/dataset/e-data-gov-il-dataset-yeshuvmoatzadata/resource/08f36575-d5f9-4c99-842c-e3516f34c31c)

לא מחקנו ושינינו שום דבר מבסיס הנתונים שלנו מכיוון שכל עמודה הייתה חשובה לצורך הניתוחים, אך היו חסרות לנו עמודות ולכן הוספנו עמודות:

עמודת שנים -הייתה עמודה שהציגה נתון תאריך מלא אשר היה מסודר לפי חודשים, אך רצינו לבצע ניתוחים לפי שנה ולכן הוספנו עמודת שנים באמצעות מניפולציה על עמודת התאריך. באופן דומה הוספנו גם עמודת חודשים.

עמודת אחוזי השמות - רצינו ליצור עמודה נוספת לניתוח הנתונים. גם עמודה זו יצרנו באמצעות מניפולציות על עמודה קיימת שנקראת השמות וחילקנו את התאים שלה בתאי העמודה שמציגה את כמות מחפשי העבודה.

חיבור בסיס הנתונים:

```
library(tidyverse)
library(ggplot2)
library(scales)
library(lubridate)
knitr::opts_chunk$set(echo = TRUE)
```

```
infodata <- read.csv("infoyeshuv1.csv",header=TRUE, sep=",", stringsAsFactors = FALSE, encodi
ng = "UTF-8") #Fetching the data

#Coloumn names corrections
names(infodata)[1] <- "Month"
names(infodata)[13] <- "Hasamot"
names(infodata)[8] <- "Women" #instead of "Wemen".

#General Additions
infodata$HasamotPerJob <-(infodata$Hasamot/infodata$Total.jobseekers) #New col. - Percentage
 of placements for job seekers.
infodata$OnlyYear <-  substr(infodata$Month,1,4)  ##New col. - extracting the year.
infodata$OnlyMonth <-  substr(infodata$Month,6,7) #New col. extracting only the month.
```

ניתוחים סטטיסטיים

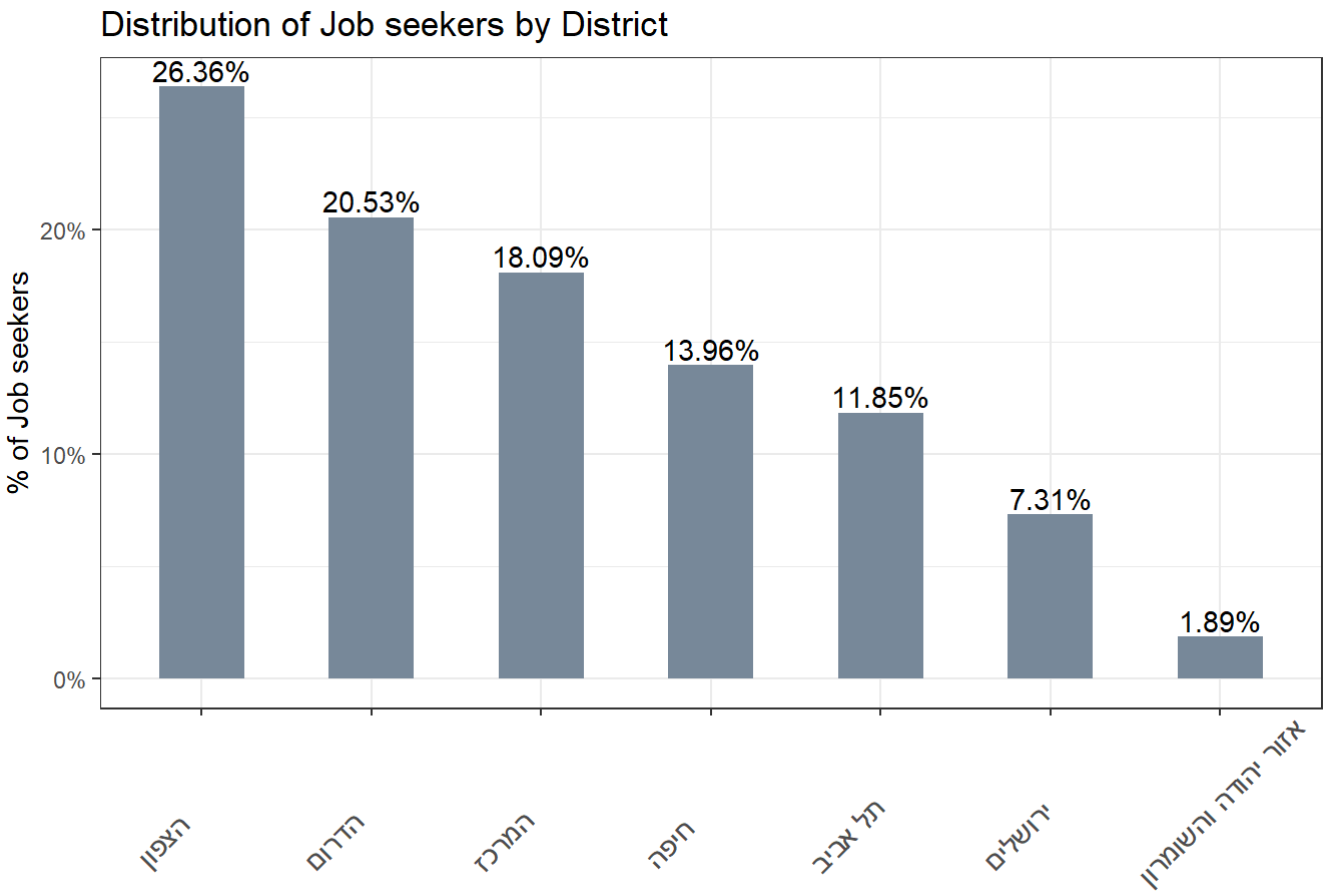
בחלק זה נציג את ההשערות שלנו לצד הניתוחים הסטטיסיטיים וכן גרפים תומכים שהפקנו מבסיס הנתונים:

האם קיימים הבדלים באחוזי המובטלים הפונים ללשכות התעסוקה בין שבעת האזורים בארץ?

```
GBarea <- aggregate(infodata[ ,"Total.jobseekers"], by=list(infodata$Cbs.district), FUN=sum)
#Summing seekers by district.
colnames(GBarea) <- c("District", "Job_seekers")
GBarea <- mutate(GBarea, Jobseek_per = GBarea$Job_seekers / sum(GBarea$Job_seekers)) #Percent
age col. of TOTAL job seekers.

#Sorting by percentage
GBarea <- GBarea[order(GBarea$Jobseek_per, decreasing = TRUE), ]
GBarea$District <- factor(GBarea$District, levels = GBarea$District)

#Visualizing the data
theme_set(theme_bw())
ggplot(GBarea, aes(y=Jobseek_per, x=District)) +
  geom_bar(position = 'dodge', stat="identity", width=.5, fill="lightslategray") +
   geom_text(aes(label=paste0(round(100*Jobseek_per,2),"%")), position=position_dodge(width=
0.1), vjust=-0.25) +
    labs(title = "Distribution of Job seekers by District" , x = "" , y = "% of Job seekers")
+
     scale_y_continuous(labels=percent) +
      theme(axis.text.x = element_text(size=11, angle=45, vjust=0.6))
```



מהגרף ניתן לראות הבדלים קיצוניים בין אזור הצפון לאזור יהודה ושומרון, לדוגמה. בחרנו לנתח הבדלים אלה באמצעות ניתוח שונות חד כיווני. מערכת ההשערות: H_0:μ_1=μ_2=μ_3=μ_4=μ_5=μ_6=μ_7

```
infodata$NewAvergeAvtala <- as.numeric(sub("%","",infodata$Estimated.unemployed.rate.in.cit
y))/100 #Convertion of Average col. into numbers.
```

```
## Warning: NAs introduced by coercion
```

```
AVERAGE_HPJ <- aggregate(infodata[,"HasamotPerJob"]*100, list(infodata$Cbs.district), mean) #
Mean of placements by district.
AVERAGE_AVTALA <- aggregate(infodata[,"NewAvergeAvtala"]*100, list(infodata$Cbs.district), me
an) #Mean of unemployement by district.
Revised_variance_HPJ <- aggregate(infodata[,"HasamotPerJob"], list(infodata$Cbs.district), va
r) #Variance of placements by district.
Count_HPJ <- aggregate(infodata[,"HasamotPerJob"], list(infodata$Cbs.district), length) #Tupl
es by district.

#ANOVA
anova.res1=aov(NewAvergeAvtala~Cbs.district, data = infodata)
summary(anova.res1)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Cbs.district     6  15.61  2.6020    1573 <2e-16 ***
## Residuals    25859  42.77  0.0017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 48 observations deleted due to missingness
```

כפי שצפינו, ערך קריטי גדול בהרבה מערך סטטיסטי, ואכן קיימים הבדלים בין האזורים באחוזי האבטלה, כפי שצפינו מראש.

ניסינו לבחון האם הפונים ללשכות תעסוקה מוגדרים כאקדמאים ולא אקדמאים. לכן מצאנו את האחוזים לכל תת מדגם כתלות באזור בממוצע השנים 2010-2019.

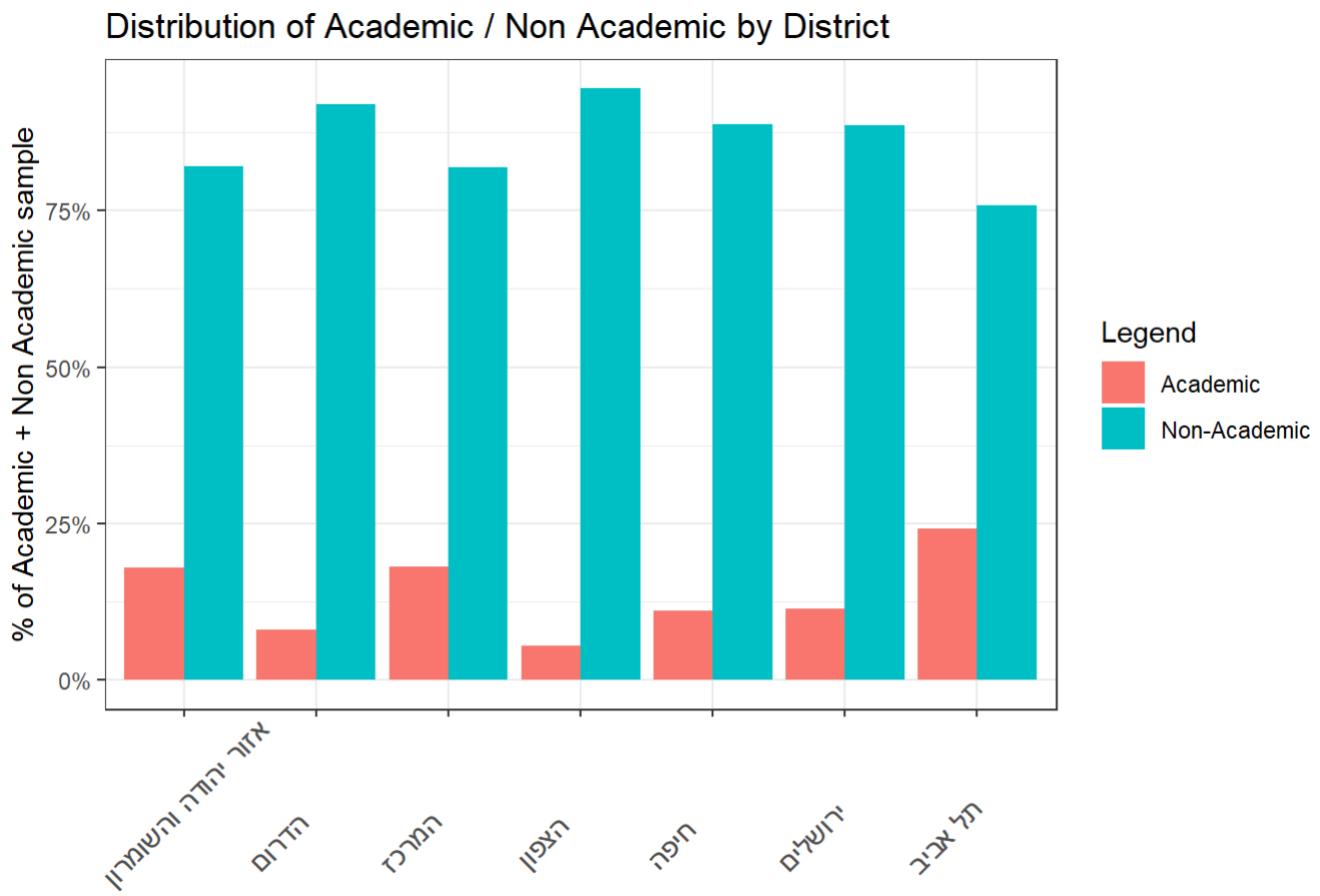
```
#Fetching the data
acadf <- data.frame(infodata$Cbs.district, infodata$Academic, infodata$Non.academic)
colnames(acadf) <- c("dis", "aca", "nonaca")

#Filters & focusing
df1 <- acadf %>%
  group_by(dis) %>%
    summarise(acaavg = mean(aca), nonacaavg = mean(nonaca)) %>%
      arrange(dis)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
compdf1 <- mutate(df1, aca_per = df1$acaavg / rowSums(df1[-1]),
                    nonaca_per = df1$nonacaavg / rowSums(df1[-1])) %>%
                       select(-acaavg, -nonacaavg) %>%
                         gather("Stat", "Value", -dis)

#Visualizing the data
ggplot(compdf1, aes(x = dis, y = Value, fill = Stat)) +
  geom_col(position = "dodge") +
      labs(title = "Distribution of Academic / Non Academic by District" , x = "" , y = "% of
Academic + Non Academic sample") +
        scale_y_continuous(labels=percent) +
          theme(axis.text.x = element_text(size=11, angle=45, vjust=0.6)) +
           scale_fill_discrete(name="Legend",
                         breaks=c("aca_per", "nonaca_per"),
                         labels=c("Academic", "Non-Academic"))
```



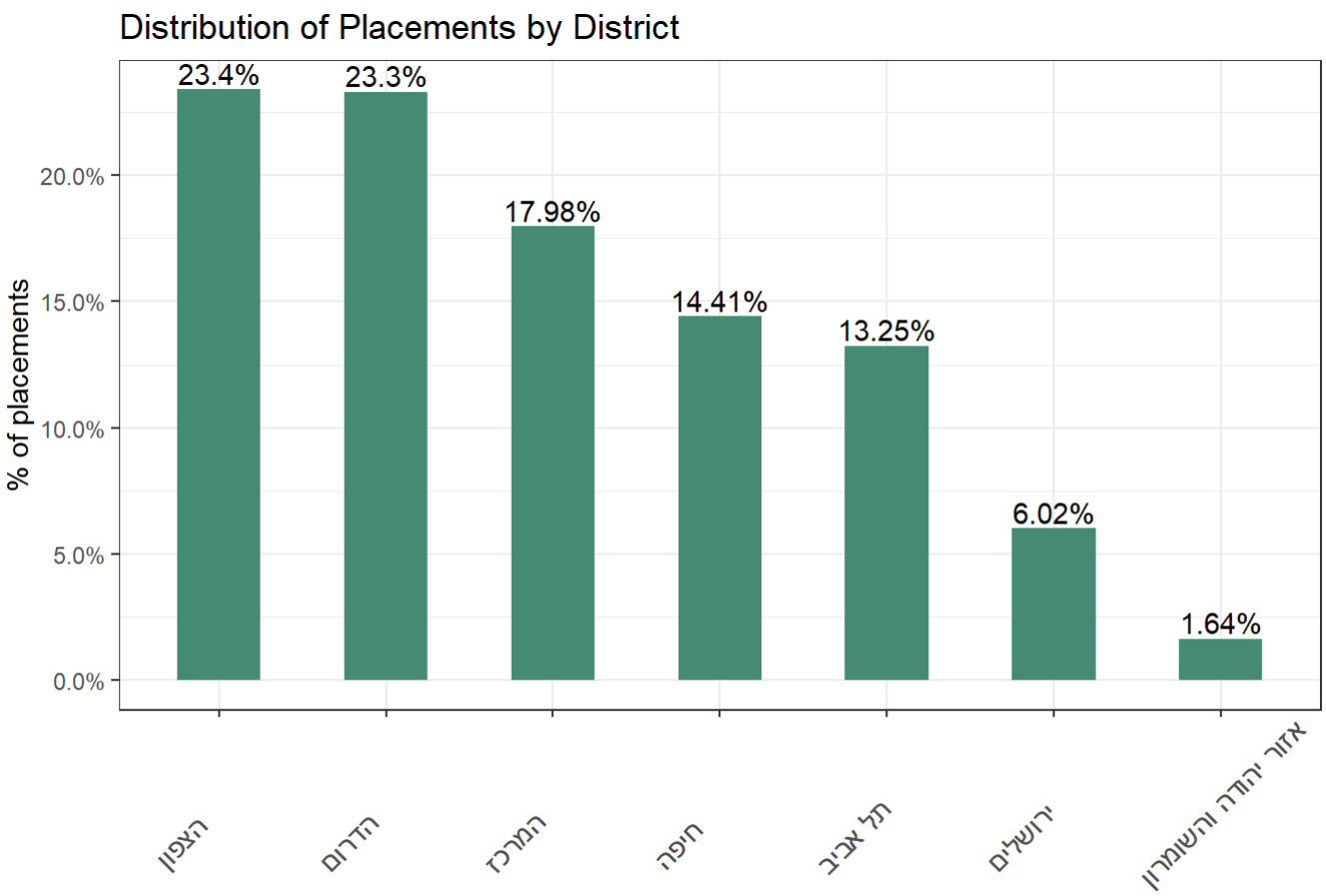
ניתן לראות שבאופף גורף, אחוז הלא אקדמאים מחפשי העבודה גדול משמעותית מאחוז האקדמאים מחפשי העבודה.

האם ישנם הבדלים באחוזי ההשמות בין שבעת האזורים בארץ?

```
#Fetching & focusing
HBarea <- aggregate(infodata[ ,"Hasamot"], by=list(infodata$Cbs.district), FUN=sum) #Summing
 placements by district.
colnames(HBarea) <- c("District", "Hasamot")
HBarea <- mutate(HBarea, placm_per = HBarea$Hasamot / sum(HBarea$Hasamot)) #Percentage col. o
f TOTAL placements.

#Sorting by percentage
HBarea <- HBarea[order(HBarea$Hasamot, decreasing = TRUE), ]
HBarea$District <- factor(HBarea$District, levels = HBarea$District)

#Visualizing the data
theme_set(theme_bw())
ggplot(HBarea, aes(y=placm_per, x=District)) +
  geom_bar(position = 'dodge', stat="identity", width=.5, fill="aquamarine4") +
   geom_text(aes(label=paste0(round(100*placm_per,2),"%")), position=position_dodge(width=0.1
), vjust=-0.25) +
    labs(title = "Distribution of Placements by District" , x = "" , y = "% of placements") +
     scale_y_continuous(labels=percent) +
      theme(axis.text.x = element_text(size=11, angle=45, vjust=0.6))
```



מהגרף ניתן לראות שיש הבדלים בכמות ההשמות לפי אזור, אך מובן שלא הסתפקנו בכך וגם כאן ביצענו ניתוח שונות חד כיווני,
כאשר מערכת ההשערות: H_0:μ_1=μ_2=μ_3=μ_4=μ_5=μ_6=μ_7

H_1:else

```
anova.res2=aov(HasamotPerJob~Cbs.district, data = infodata)
summary(anova.res2)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Cbs.district     6   0.91  0.1514   89.04 <2e-16 ***
## Residuals    25907  44.04  0.0017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

גם כאן נקבל את השערתנו שהוכחה באמצעות ניתוח השונות ונצהיר שאחוזי ההשמה שונים בין האזורים.

האם יש הבדל באחוז ההשמות של לשכות התעסוקה לאורך השנים?

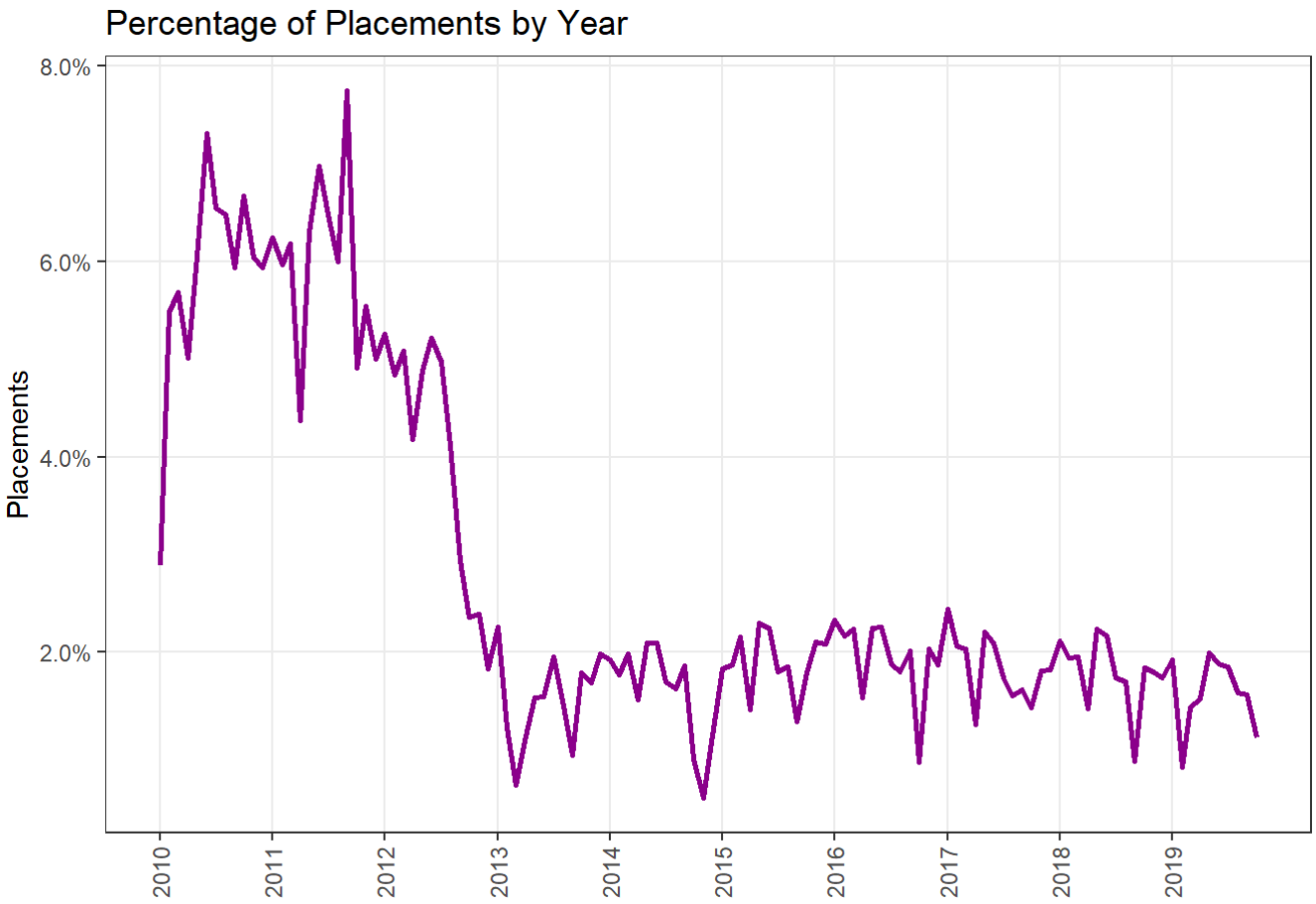
רצינו לנתח את אחוזי ההשמות לאורך השנים. לצורך כך התחלנו בגרף המציג ממוצע השמות ארצי לכל שנה:

```
#Fetching and re-labeling the data
HasamotPY <- aggregate(infodata[,"HasamotPerJob"], list(infodata$Month), FUN = mean)
colnames(HasamotPY) = c("date", "hasamot")

HasamotPY$date <- parse_date_time(HasamotPY$date, orders = c("Ym", "mY"))
HasamotPY$date <- as.Date(HasamotPY$date, format="%Y-%m-%d")

#For yearly x-axis tags.
brks <- HasamotPY$date[seq(1, length(HasamotPY$date), 12)]
lbls <- lubridate::year(brks)

#Visualizing the data
theme_set(theme_bw())
ggplot(HasamotPY, aes(x=date, y=hasamot)) +
  geom_line(size=1, color="darkmagenta") +
    labs(title="Percentage of Placements by Year", x="", y="Placements") +
     scale_x_date(labels = lbls, breaks = brks) +
         scale_y_continuous(labels=percent) +
          theme(axis.text.x = element_text(angle = 90, vjust=0.5), panel.grid.minor = element
_blank())
```



רצינו לבדוק את ההשערה שאחוז ההשמות בשנים 2010-2012 גדול מאחוז ההשמות בשנים 2013-2019. לצורך כך, ובגלל שהשונות בתחום זה לא ידועה לנו, השתמשנו במבחן t.

מיצענו את כלל אחוזי ההשמות לפי 2 קבוצות מדגמים: {1 – 2010-2012, 2- 2013-2019 }, הנחנו כי רמת המובהקות הינה 5%.
מערכת ההשערות כדלקמן:

H0: p1 – p2 = 0

H1: p1 – p2 > 0

```
mu0 <- mean(infodata$HasamotPerJob)

p1<-subset(infodata, OnlyYear >= 2010 & OnlyYear <= 2012 ,
           select=c("HasamotPerJob", "OnlyYear"))[,"HasamotPerJob"]

p2<-subset(infodata, OnlyYear >=2013,
           select=c("HasamotPerJob", "OnlyYear"))[,"HasamotPerJob"]

t.test(p1-p2, NULL, alternative="greater", mu=mu0,
       var.equal = FALSE, conf.level = 0.95)
```

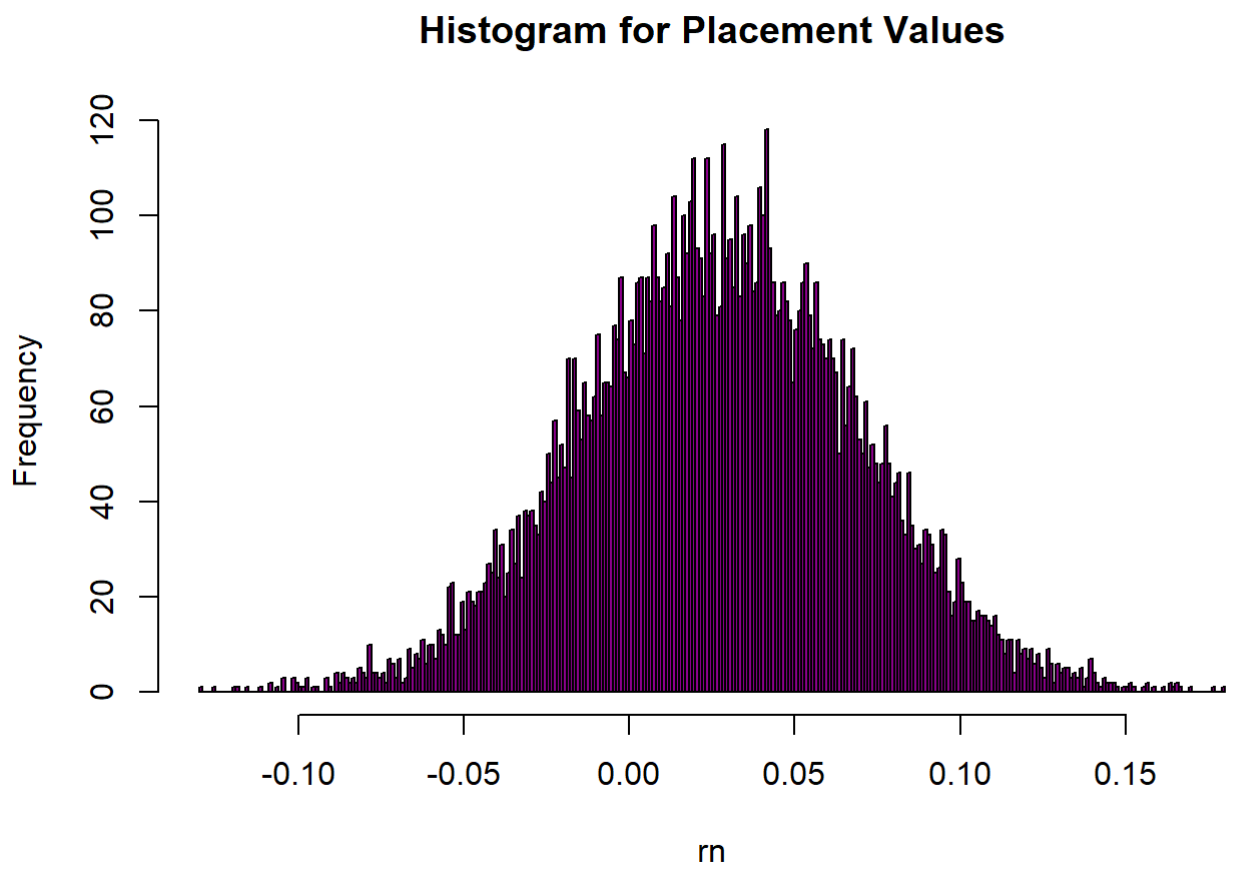
```
## Warning in p1 - p2: longer object length is not a multiple of shorter object
## length
```

```
##
##  One Sample t-test
##
## data:  p1 - p2
## t = 23.659, df = 18893, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0.02704109
## 95 percent confidence interval:
##  0.03612356        Inf
## sample estimates:
##  mean of x
## 0.03680223
```

מהפלט המתקבל הסקנו שהשערתינו נכונה, ערך P-value שואף לאפס. נשים לב שגם קיבלנו אזהרה על כך ששני המדגמים לא באותו גודל, או לפחות לא מתחלקים.

בהמשך לזאת רצינו לבדוק האם אחוזי ההשמה מתפלגים נורמלית. לשם כך ניסינו תחילה לחזות את התוצאה באמצעות היסטוגרמה:

```
rn<-rnorm(10000,mu0, sd=sd(infodata$HasamotPerJob))
hist(rn, main="Histogram for Placement Values",col="darkmagenta",breaks = 400)
```



ניתן לראות שהצורה המתקבלת קרובה לצורת הפעמון אך לא באופן סימטרי לחלוטין. נמשיך לבדיקת התפלגות נורמלית קולמוגורוב-סמירנוף עבור מדגם בודד.

```
ks.test(x=as.numeric(infodata$HasamotPerJob), pnorm)
```

```
## Warning in ks.test(x = as.numeric(infodata$HasamotPerJob), pnorm): ties should
## not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  as.numeric(infodata$HasamotPerJob)
## D = 0.5, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

ערך הסטטיסטי שקיבלנו (0.5), אינו אופטימלי להחלטה כיוון שלא שואף ל0 או ל1. הנתון תואם לפלט ההיסטוגרמה. היות וערך pValue שואף ל0, נוכל להסיק כי ייתכן ואחוזי ההשמה מתפלגים בצורה נורמלית כלשהי, אך לא בהכרח תחת התפלגות נורמלית סטנדרטית.

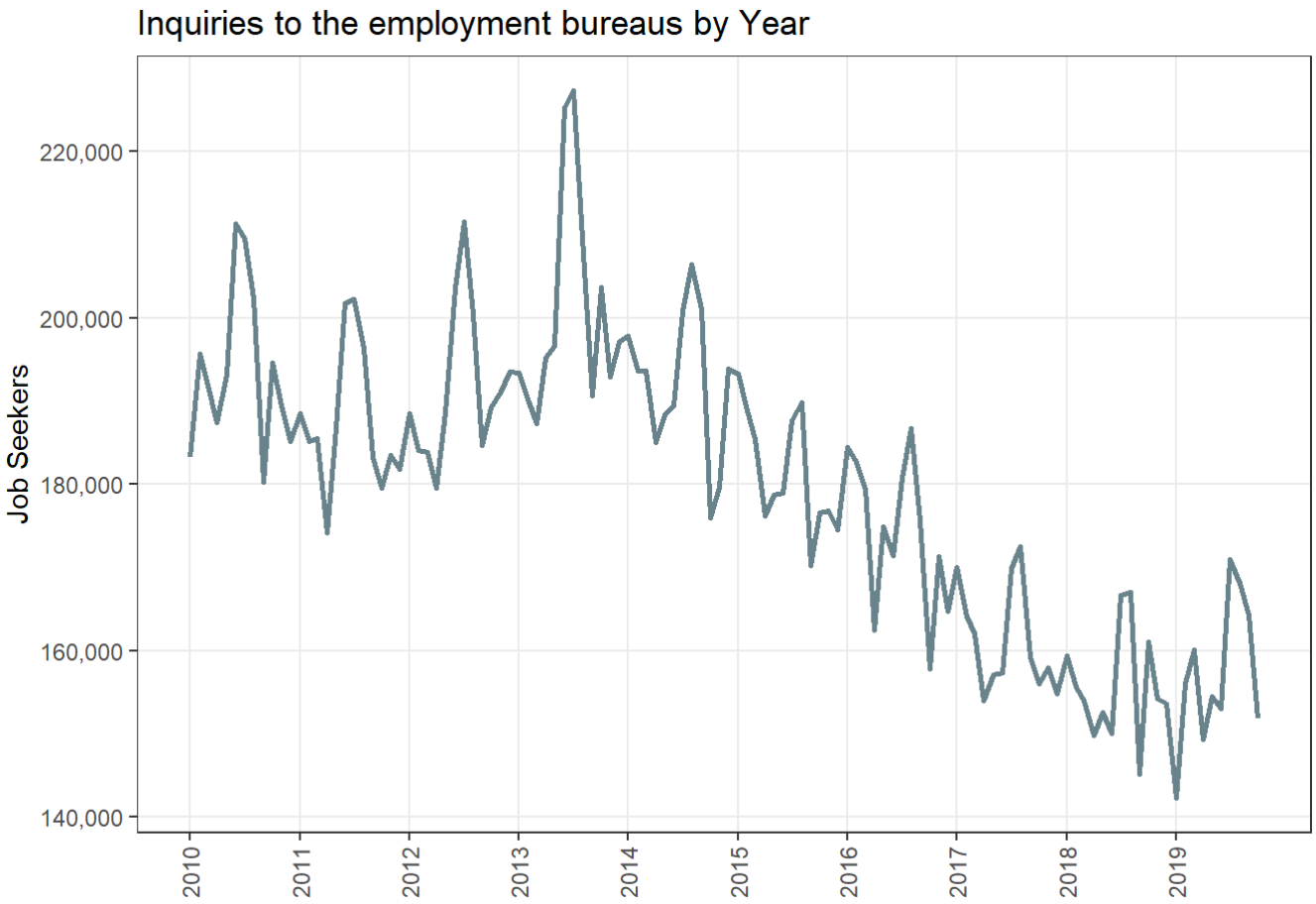
נתונים כללים שבחרנו להציג על מנת להראות את יעילות הלשכות

```
#Fetching and re-labeling the data
tspy <- aggregate(infodata[,"Total.jobseekers"],  list(infodata$Month), sum)
colnames(tspy) = c("date", "total_job_s")

tspy$date <- parse_date_time(tspy$date, orders = c("Ym", "mY"))
tspy$date <- as.Date(tspy$date, format="%Y-%m-%d")

#For yearly x-axis tags.
brks <- tspy$date[seq(1, length(tspy$date), 12)]
lbls <- lubridate::year(brks)

#Visualizing the data
theme_set(theme_bw())
ggplot(tspy, aes(x=date, y=total_job_s)) +
  geom_line(size=1, color="lightblue4") +
    labs(title="Inquiries to the employment bureaus by Year", x="", y="Job Seekers") +
     scale_x_date(labels = lbls, breaks = brks) +
         scale_y_continuous(labels=comma) +
          theme(axis.text.x = element_text(angle = 90, vjust=0.5), panel.grid.minor = element
_blank())
```



ניתן לראות את הירידה בכמות הפונים ללשכות משנת 2013 ומעלה.

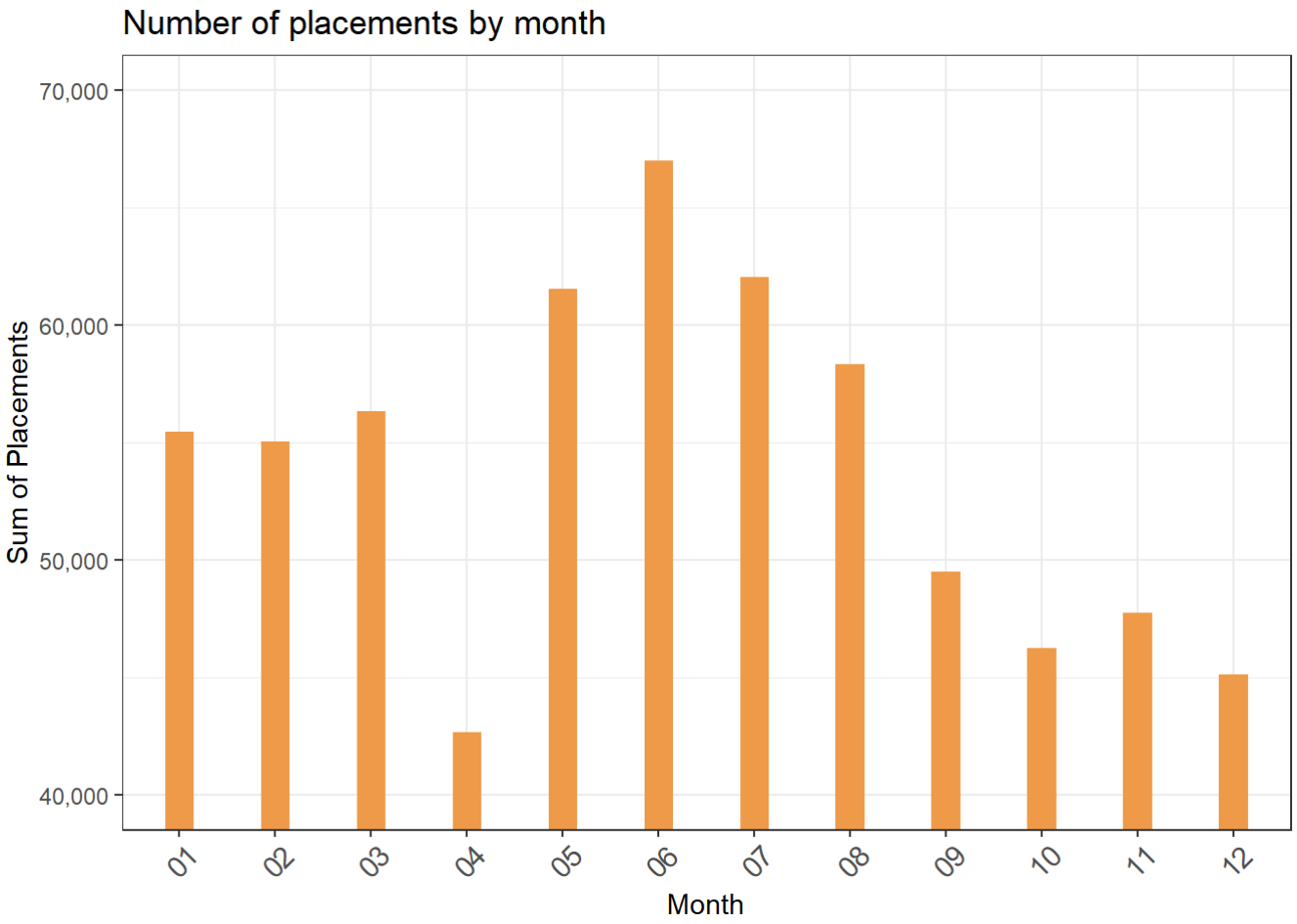
```
tspm <- aggregate(infodata[,"Total.jobseekers"], list(infodata$OnlyMonth), sum)
colnames(tspm) <- c("month", "total_job_sm")


#Visualizing the data
theme_set(theme_bw())
ggplot(tspm, aes(y=total_job_sm, x=month)) +
  geom_bar(stat="identity", width=.3, fill="deeppink3") +
    labs(title = "Inquiries to the employment bureaus by month" , x = "Month" , y = "Sum of J
ob seekers") +
     scale_y_continuous(labels=comma) +
      theme(axis.text.x = element_text(size=11, angle=45, vjust=0.6)) +
        coord_cartesian(ylim=c(1500000, 2000000)) #Zooming on y = Job Seekers values(!)
```

בגרף זה ניתן לראות את כמות הפניות ללשכות לפי חודשים, ניכר שבחודשי הקיץ כמות הפניות גדלה יחסית לשאר השנה. נשתמש בניתוח זה על מנת לסייע לנו בהסקת המסקנות.

```
#Fetching and renaming the data
hspm<- aggregate(infodata[,"Hasamot"], list(infodata$OnlyMonth), sum) #סידור כמות ההשמות לפי חודש
colnames(hspm) <- c("month", "hasamot")

#Visualizing the data
theme_set(theme_bw())
ggplot(hspm, aes(y=hasamot, x=month)) +
  geom_bar(stat="identity", width=.3, fill="tan2") +
    labs(title = "Number of placements by month" , x = "Month" , y = "Sum of Placements") +
     scale_y_continuous(labels=comma) +
      theme(axis.text.x = element_text(size=11, angle=45, vjust=0.6)) +
        coord_cartesian(ylim=c(40000, 70000)) #Zooming on y = Placements values(!)
```



אם נבחן את כמות ההשמות, נשים לב שכמות ההשמות הקטנה ביותר היא בודש אפריל. כמות ההשמות הגדולה ביותר - בחודש יוני.

סיכום ומסקנות

הראנו שכמות מבקשי העבודה הפונים ללשכת התעסוקה יורדת עם השנים, ושבחודשים מסויימים כמות ההשמות קטנה מאוד עד שואפת לאפס. עם כל זאת, הנתונים במדינה מראים על שיעור אבטלה נמוך מאי פעם. מתוך כך אנו יכולים להסיק שהגורם לשגשוג בשוק העבודה הוא אינו לשכת התעסוקה, ושהיא אינה רלוונטית כפי שהייתה בעבר. אנו יכולים רק לשער שהגורם לכך הוא ככל הנראה הקדמה הטכנולוגית, שמאפשרת לאנשים למצוא עבודה בקלות מבלי צורך להסתמך על לשכת התעסוקה, וכמו כן יוקר המחיה אשר מניע אנשים רבים לצאת לעבודה ולא להיות בבטלה.

חשוב לציין כי התקציב המוקצה ללשכות התעסוקה במדינה עמד בשנת 2019 על 278,355,215 ₪. זו אומנם אכן ירידה של 2% משנת 2018, אך עדיין סכום לא מבוטל במדינתנו הקטנה. מכאן שהיינו ממליצים להפעיל את לשכות התעסוקה באזורים בהם היא יותר רלוונטית כמו הצפון והדרום, ולבטל את פועלה באזורי יהודה ושומרון. כמו כן היינו ממליצים להפעיל את לשכות התעסוקה רק בחודשי הקיץ בהם ראינו שהלשכות פועלות יותר, על מנת לחסוך בתקציב.