Placeholder Golf Project

Thomas Schechter and Ian Bogley

12/10/2021

Read data in and installed packages.

```
#Package unloading
library(pacman)
p_load(tidyverse,data.table,tidymodels,stargazer,ggpubr,janitor,sandwich)

#Read in data
pga_data17 <- fread("PGATOUR_data2.csv")
```

Subset the data to remove null and na values, turned subset data into data.table

```
#Subset data
pga_data <- pga_data17[1:195,] %>% clean_names() %>%
  mutate(top_ten_finisher = (number_of_top_tens>0))
```

Now, let's graph potential factors in top ten finishes and look for correlations.

```
#Begin exploring the relationships in the data

#Correlation b/w rounds played and top tens
model1 <- lm(rounds_played~number_of_top_tens, pga_data)

tidy(model1)
```

```
## # A tibble: 2 x 5
##   term               estimate std.error statistic   p.value
##   <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           74.0      1.53      48.5  5.33e-110
## 2 number_of_top_tens     1.65     0.443      3.72 2.60e-  4
```

```
#Plot the rounds played relationship
round_plot <-  ggplot(pga_data, aes(rounds_played, number_of_top_tens)) +
  geom_point() +  geom_smooth(method = "lm") +
  labs(x = "Rounds Played",
       y = "Number of Top Tens")

#Plotting strokes gained putting vs. top ten finishes
putt_plot <- ggplot(pga_data, aes(sg_putting_per_round, number_of_top_tens)) +
  geom_point() +  geom_smooth(method = "lm") +
  labs(x = "Shots Gained Putting per Round",
```

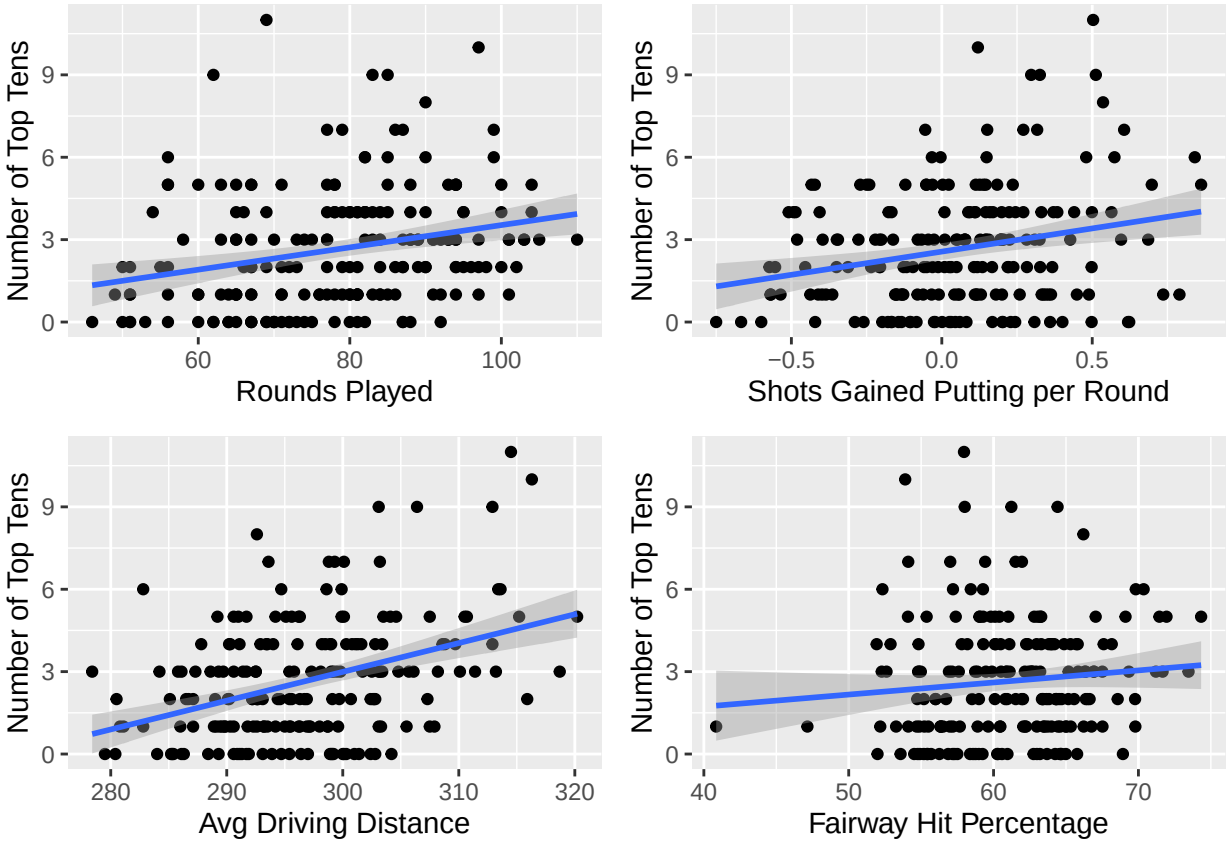
```
       y = "Number of Top Tens")

#Plotting average drive distance vs. top ten finishes
drive_plot <- ggplot(pga_data, aes(avg_driving_distance,number_of_top_tens)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(x = "Avg Driving Distance",
       y = "Number of Top Tens")

#Plotting Fairway hit percentage vs top ten finishes
fairway_plot <- ggplot(pga_data, aes(x=fairway_hit_percent,y = number_of_top_tens)) +
  geom_point() + geom_smooth(method = "lm")+
  labs(x = "Fairway Hit Percentage",
       y = "Number of Top Tens")

#Arrange previous plots into a single graphic
ggarrange(round_plot,putt_plot,drive_plot,fairway_plot)
```

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



Not a significant amount of visual correlation between number of rounds played and number of top ten finishes.

Strokes gained putting seems to have more of an effect on success.

Some relationship appears present in average driving distance, though with more outliers on both ends of the spectrum.

There seems to be even less of a correlation in terms of fairway hit percentage than the previous ones. Even if there is, it is much smaller of an impact than the other factors examined.

Let's examine these relationships as linear regressions, identifying the impact of each and the robustness of the coefficients.

```
for (i in 1:4) {
  #vector of variable names
  variables <- c("rounds_played","sg_putting_per_round",
                 "avg_driving_distance","fairway_hit_percent")

  #models for each variable
  eval(parse(text = paste("model_",i+1,
                          " <- lm(number_of_top_tens ~ ",variables[i],
                          ",pga_data)",sep = "")))

  #robust se for each variable
  eval(parse(text = paste("robust_se_",i+1," <- list(sqrt(diag(vcovHC(model_",i+1,"))))",sep ="")))
}
```

```
stargazer(model_2,model_3,model_4,model_5)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Dec 15, 2021 - 4:13:41 PM

Table 1:

	Dependent variable:			
	number_of_top_tens			
	(1)	(2)	(3)	(4)
rounds_played	0.041***			
	(0.011)			
sg_putting_per_round		1.689***		
		(0.490)		
avg_driving_distance			0.105***	
			(0.018)	
fairway_hit_percent				0.044
				(0.031)
Constant	−0.528	2.564***	−28.463***	−0.036
	(0.868)	(0.155)	(5.288)	(1.901)
Observations	195	195	195	195
R^2	0.067	0.058	0.152	0.010
Adjusted R^2	0.062	0.053	0.148	0.005
Residual Std. Error (df = 193)	2.132	2.142	2.032	2.196
F Statistic (df = 1; 193)	13.847***	11.874***	34.642***	2.011
Note:				*p<0.1; **p<0.05; ***p<0.01