Blake Pappas

5 December 2022

NASCAR Racing: More Than Just Driving Fast and Turning Left?

Figure 1: Watkins Glen International Road Course

INTRODUCTION

The National Association for Stock Car Auto Racing (NASCAR) is the largest auto racing sanctioning body in the United States. Each year, NASCAR organizes hundreds of races across 48 states, as well as in Canada, Mexico, and Europe. Known worldwide for its colorful paint schemes, banked turns, high speeds, and aggressive racing style, the sport has evolved greatly throughout its 74-year history.

In recent years, NASCAR has imposed new rules and regulations regarding how teams may build and operate their cars, so much so that many critics of the sport have argued that it's now easier than ever to be successful. Can driver success be defined by more than merely going fast and turning left?

The following analysis examines data collected from the 2021 NASCAR Cup Series race at the Watkins Glen International road course. The goal of this analysis is to determine which factors go about best explaining success on the race track. For the context of this analysis, success will be evaluated by finishing position and points earned. The lower the finishing position, the better the driver performed. The more points earned, the better the driver performed.

DATA

The data used in this analysis originated from nascar.com. It was collected on Sunday, August 8, 2021. NASCAR's original purpose in collecting this data was to build an HTML tool to provide a live, side-by-side comparison of two drivers throughout the race. This tool reports on critical race statistics like starting position, lap time, lap speed, highest running position, lowest running position, and current running position.

DATA IMPORTATION

Contrary to much of the data for professional sporting events, the data for this analysis was not readily available. Because the data was used to create a highly interactive HTML tool, NASCAR did not conveniently advertise this kind of information. In order to retrieve the data needed for this analysis, a web scrape was necessary. This required the underlying HTML of nascar.com to be inspected and parsed until all appropriate JSON files containing the data for the 2021 Watkins Glen race could be found. Once these JSON files were discovered and retrieved, they were converted into the following CSVs:

- race.csv
- team.csv
- sponsor.csv
- points.csv
- manufacturer.csv
- driver.csv

These CSVs were then imported to the SAS server as follows:

```
PROC IMPORT DATAFILE="/home/wpappas/DSA 8030/SAS Project/driver.csv"
            DBMS=CSV OUT=SAS_PROJ.DRIVER REPLACE;
RUN;

PROC IMPORT DATAFILE="/home/wpappas/DSA 8030/SAS Project/manufacturer.csv"
            DBMS=CSV OUT=SAS_PROJ.MANUFACTURER REPLACE;
RUN;

PROC IMPORT DATAFILE="/home/wpappas/DSA 8030/SAS Project/points.csv"
            DBMS=CSV OUT=SAS_PROJ.POINTS REPLACE;
RUN;

PROC IMPORT DATAFILE="/home/wpappas/DSA 8030/SAS Project/race.csv"
            DBMS=CSV OUT=SAS_PROJ.RACE REPLACE;
RUN;

PROC IMPORT DATAFILE="/home/wpappas/DSA 8030/SAS Project/sponsor.csv"
            DBMS=CSV OUT=SAS_PROJ.SPONSOR REPLACE;
RUN;

PROC IMPORT DATAFILE="/home/wpappas/DSA 8030/SAS Project/team.csv"
            DBMS=CSV OUT=SAS_PROJ.TEAM REPLACE;
```

Figure 2: Data Importation Through *PROC IMPORT*

TIDYING THE DATA

In general, the raw, underlying data for this analysis was very clean. Few steps were needed to tidy the data into a reporting-level quality. The first step taken to tidy the data was to join all six CSVs together to create a master data set. The base CSV for this master data set was race.csv. The other five CSVs were consecutively left joined (using a PROC SQL statement) to race.csv to create the WATKINS_GLEN table (located in the SAS_PROJ library). All CSVs were joined together by the *car_number* field, which was the unique identifier (primary key) in all six individual raw data sets.

The columns of the WATKINS_GLEN table were reordered in the SELECT clause of the PROC SQL statement in order for the table to follow a better defined and structured format. The columns were ordered in such a way as to allow for more qualitative data fields to be presented first (i.e. car_number, driver, manufacturer), followed by more quantitative data fields (i.e. lap_time, lap_speed, points).

```
PROC SQL;
CREATE TABLE SAS_PROJ.WATKINS_GLEN AS
SELECT
          RACE.car_number
         ,DRIVER.driver
         ,MANUFACTURER.manufacturer
         ,TEAM.team
         ,SPONSOR.sponsor
         ,RACE.start
         ,RACE.finish
         ,POINTS.points
         ,RACE.lap
         ,RACE.lap_time
         ,RACE.lap_speed
         ,RACE.running_position
FROM SAS_PROJ.RACE AS RACE
LEFT JOIN SAS_PROJ.TEAM AS TEAM
    ON RACE.car_number = TEAM.car_number
LEFT JOIN SAS_PROJ.SPONSOR AS SPONSOR
    ON RACE.car_number = SPONSOR.car_number
LEFT JOIN SAS_PROJ.POINTS AS POINTS
    ON RACE.car_number = POINTS.car_number
LEFT JOIN SAS_PROJ.MANUFACTURER AS MANUFACTURER
    ON RACE.car_number = MANUFACTURER.car_number
LEFT JOIN SAS_PROJ.DRIVER AS DRIVER
    ON RACE.car_number = DRIVER.car_number
ORDER BY RACE.finish, RACE.lap;
QUIT;
```

Figure 3: Tidying the Data Through *PROC SQL*

CONDENSED MASTER DATA TABLE

In all, the finished data set contained 12 variables and 3,330 observations. This data set contained 138 instances of missing values, coming from only the *lap_time* and *lap_speed* variables. In order to keep the integrity of the data intact, missing values were not transformed in any way.

	car_number	driver	manufacturer	team	sponsor	start	finish	points	lap	lap_time	lap_speed	running_position
1	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	1	76.753	114.914	4
2	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	2	73.503	119.995	3
3	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	3	73.377	120.201	3
4	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	4	73.26	120.393	3
5	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	5	73.502	119.997	3
6	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	6	73.681	119.705	3
7	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	7	74.546	118.316	3
8	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	8	74.067	119.081	3
9	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	9	74.446	118.475	3
10	5	Kyle Larson	Chevrolet	Hendrick Motorsports	HendrickCars.com	4	1	56	10	74.904	117.751	2

Figure 4: WATKINS_GLEN Table

What exactly does the data in this table mean?

Each row of the WATKINS_GLEN provides an in-depth summary of any given lap for a particular driver. For example, consider row 1 from the data table, which has the following observations:

- ***car_number*:** 5
- ***driver*:** Kyle Larson
- ***manufacturer*:** Chevrolet
- ***team*:** Hendrick Motorsports
- ***sponsor*:** HendrickCars.com
- ***start*:** 4
- ***finish*:** 1
- ***points*:** 56
- ***lap*:** 1
- ***lap_time*:** 76.753
- ***lap_speed*:** 114.914
- ***running_position*:** 4

This row represents the lap 1 results for Kyle Larson, the driver of the #5 Hendrick Motorsports HendrickCars.com Chevrolet. This lap took him 76.753 seconds to complete, equaling an overall speed of approximately 114.914 miles per hour. As he crossed the start/finish line to complete this first lap, Larson was in 4th place.

DATA ANALYSIS

VARIABLE DESCRIPTIONS, STATISTICS, AND DISTRIBUTIONS

car_number

Description: The driver's car number
Data Type: numeric
Minimum: 0
1st Quartile: 9
Median: 20
Mean: 27.32
3rd Quartile: 42
Maximum: 99

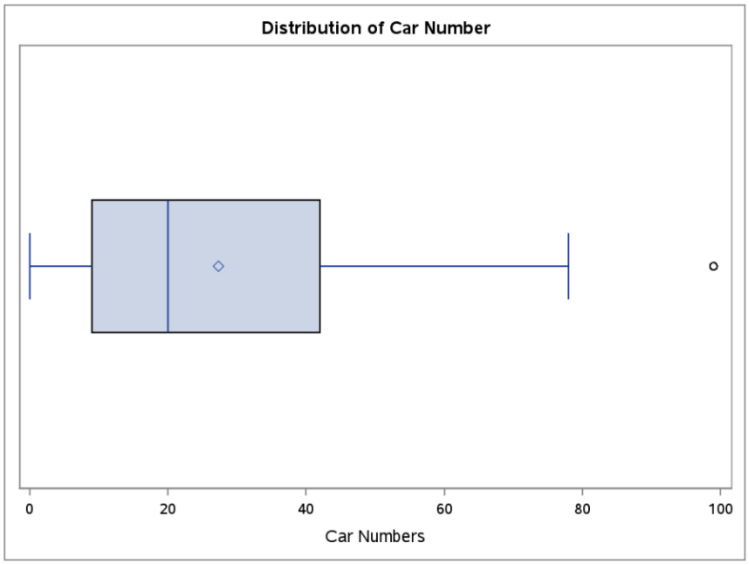

Figure 5: *car_number* Boxplot

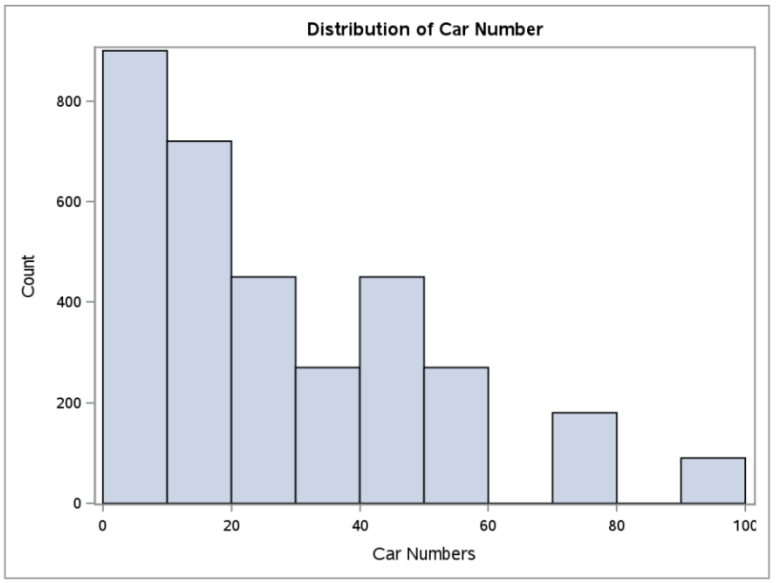

Figure 6: *car_number* Histogram

The *car_number* variable is a unique identifier representing the number on the car which the driver used during the race. Looking at this variable's boxplot and histogram, it is evident that *car_number* is positively-skewed, as 31 drivers had car numbers below 50, compared to only 6 drivers having car numbers above 50.

driver

Description: The first and last name of the driver
Data Type: character
Number of Unique Drivers: 37
Drivers: Kyle Larson, Chase Elliott, Martin Truex Jr., Kyle Busch, Denny Hamlin, William Byron, Christopher Bell, Kevin Harvick, Chase Briscoe, Tyler Reddick, Matt DiBenedetto, Ross Chastain, Kurt Busch, Ryan Blaney, Austin Dillon, Aric Almirola, Chris Buescher, Cole Custer, Ricky Stenhouse Jr., Alex Bowman, Michael McDowell, Joey Logano, Bubba Wallace, Corey LaJoie, Ryan Newman, Anthony Alfredo, Erik Jones, Ryan Preece, Justin Haley, Kyle Tilley, Daniel Suarez, Quin Houff, Josh Bilicki, RC Enerson, Brad Keselowski, Garrett Smithley, James Davison

manufacturer

Description: The make of the car
Data Type: character
Number of Unique Manufacturers: 3
Manufacturers: Chevrolet, Toyota, Ford

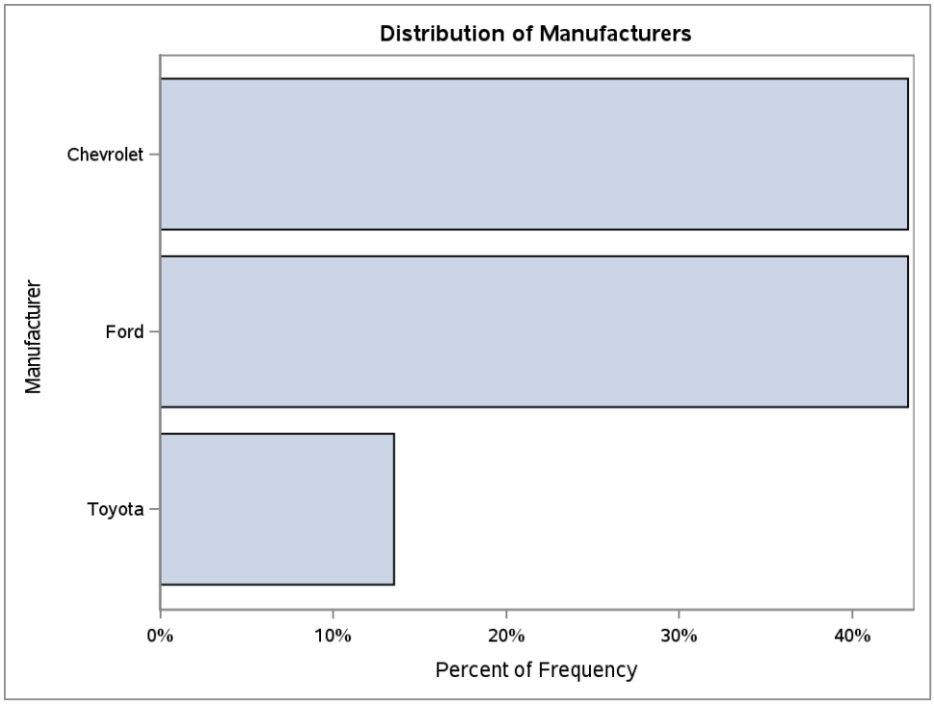

Figure 7: *manufacturer* Frequency Bar Graph

The *manufacturer* variable represents the make of the car being raced. According to the bar chart, only three manufacturers participated in the race: Chevrolet, Toyota, and Ford. The chart also indicates that Chevrolet and Ford were equally matched, with each having 16 cars in the race. Toyota, on the other hand, was the minority, having just 5 cars compete in the race.

team

Description: The motorsport team to which the driver belongs
Data Type: character
Number of Unique Teams: 19
Teams: Hendrick Motorsports, Joe Gibbs Racing, Stewart-Haas Racing, Richard Childress Racing, Wood Brothers Racing, Chip Ganassi Racing, Team Penske, Roush-Fenway Racing, JTG-Daugherty Racing, Front Row Motorsports, 23XI Racing, Spire Motorsports, Richard Petty Motorsports, Live Fast Motorsports, Trackhouse Racing, StarCom Racing, Rick Ware Racing, Skip Barber Racing, Petty Ware Racing

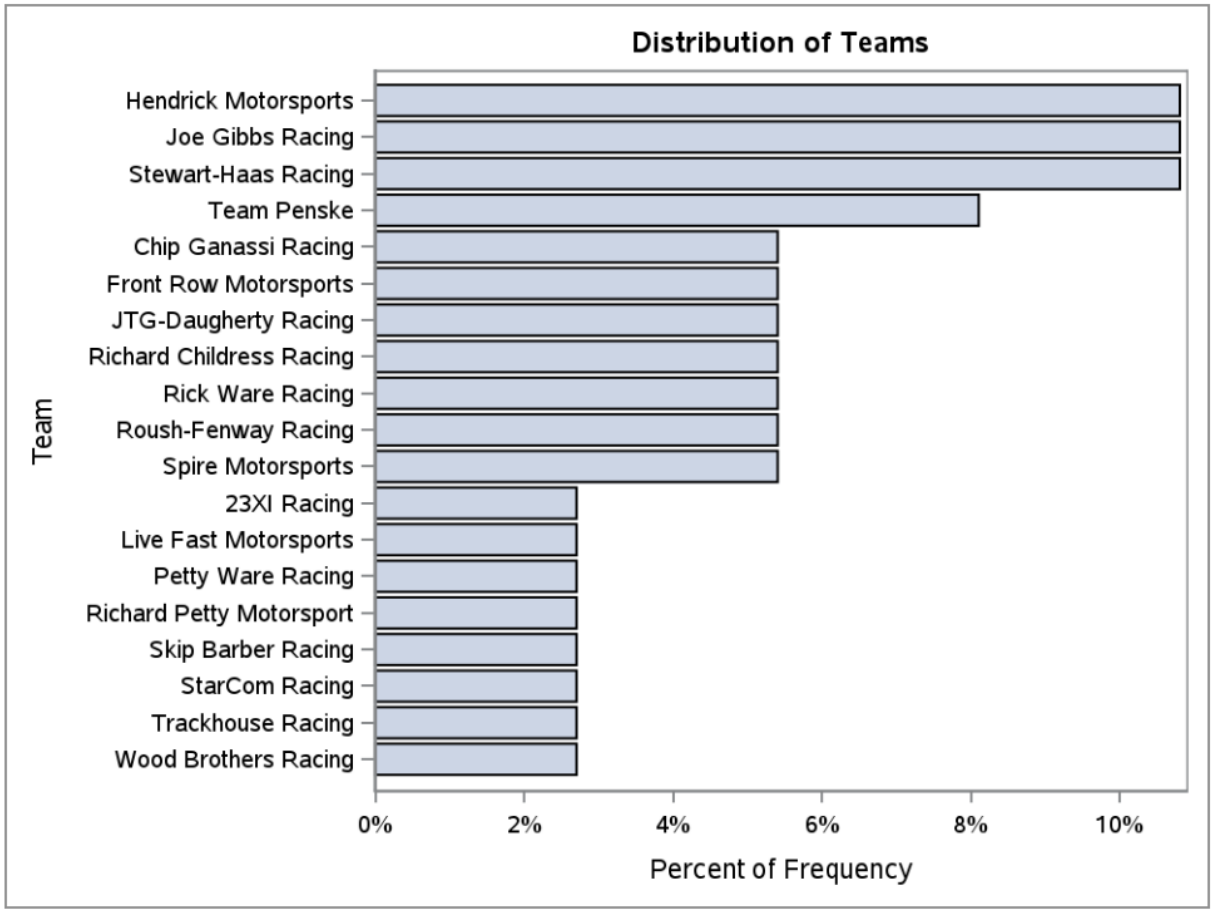

Figure 8: *team* Frequency Bar Graph

The *team* variable represents the motorsport team that participated in the race. According to the bar chart, there were 19 teams that competed in the race. The number of cars per team ranged from 1 to 4. Interestingly enough, three teams (Hendrick Motorsports, Joe Gibbs Racing, and Stewart-Haas Racing) had a combined 12 cars participate, constituting nearly 33% of all cars in the race.

sponsor

Description: The driver's primary sponsor(s) of the driver's race car
Data Type: character
Number of Unique Sponsors: 36
Sponsors: HendrickCars.com, NAPA Auto Parts, Reser's Fine Foods, Snickers, FedEx Express, Axalta, STANLEY, Busch Light Apple, HighPoint.com, Chevrolet Accessories, Menards/Moen, MyMcDonald's Rewards, Monster Energy, DEX Imaging, Cowboy Channel, Go Bowling, socios.com, HaasTooling.com, Kroger/Bush's Beans, Ally, CarParts.com, Verizon 5G, Toyota, Nations Guard, Bence Motor Sales, Clean Harbors, Kleenex, Fraternal Order of Eagles, Bremont Chronometers, Good Sam, Fare/Share, Insurance King, Lucas Oil School of Racing, Wabash National, Skip Barber Racing, Nurtec ODT

start

Description: The position where the driver started the race
Data Type: numeric
Minimum: 1
1st Quartile: 10
Median: 19
Mean: 19
3rd Quartile: 28
Maximum: 37

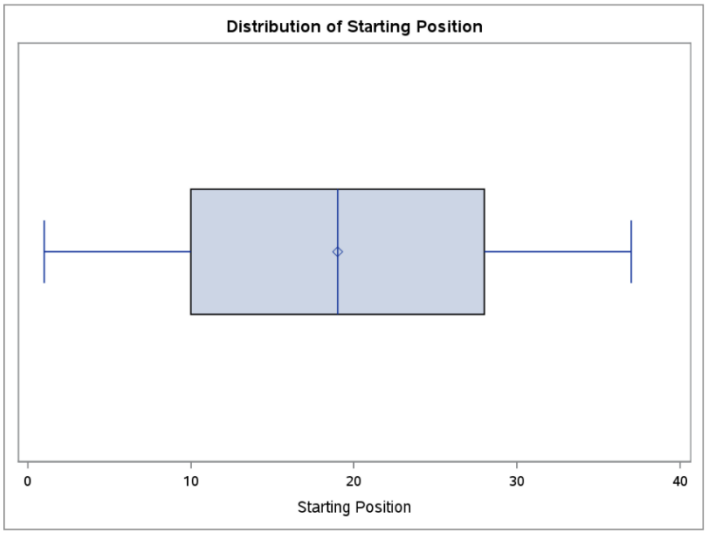

Figure 9: *start* Boxplot

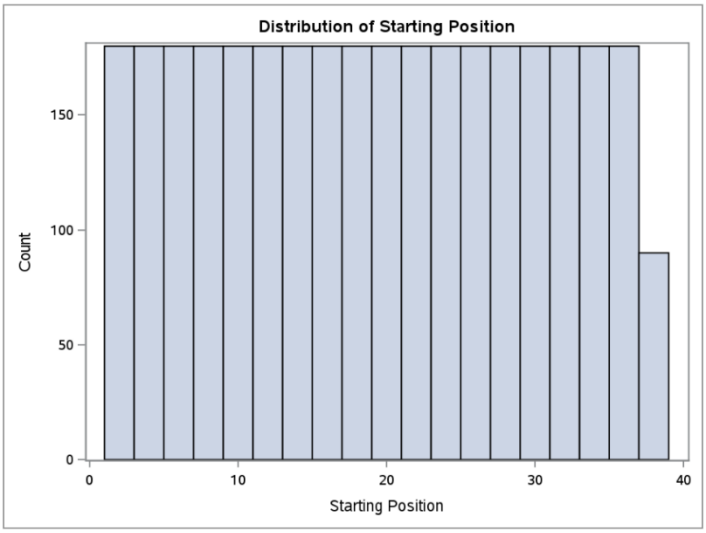

Figure 10: *start* Histogram

The *start* variable represents the driver's starting position in the race. The boxplot and histogram indicate that this variable is unimodal, symmetric, and normally distributed. This should not be of much surprise, as each starting position can only be unique to one driver.

finish

Description: The position where the driver finished the race
Data Type: numeric
Minimum: 1
1st Quartile: 10
Median: 19
Mean: 19
3rd Quartile: 28
Maximum: 37

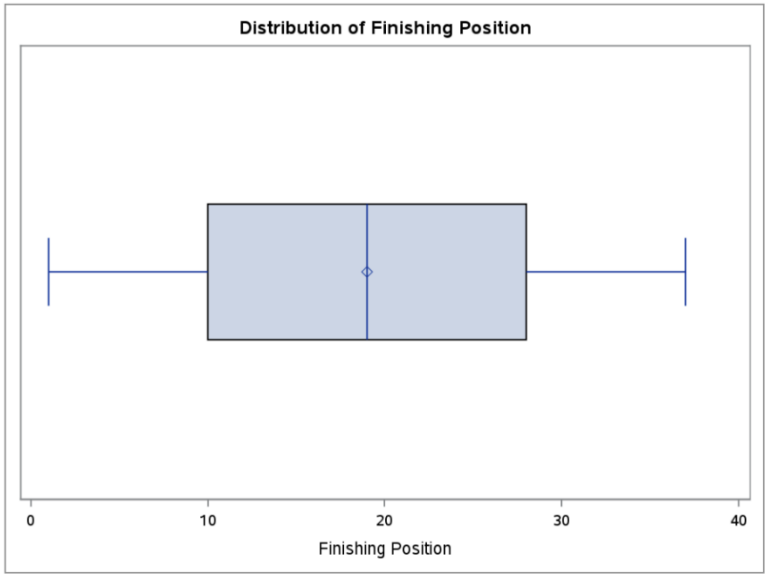

Figure 11: *finish* Boxplot

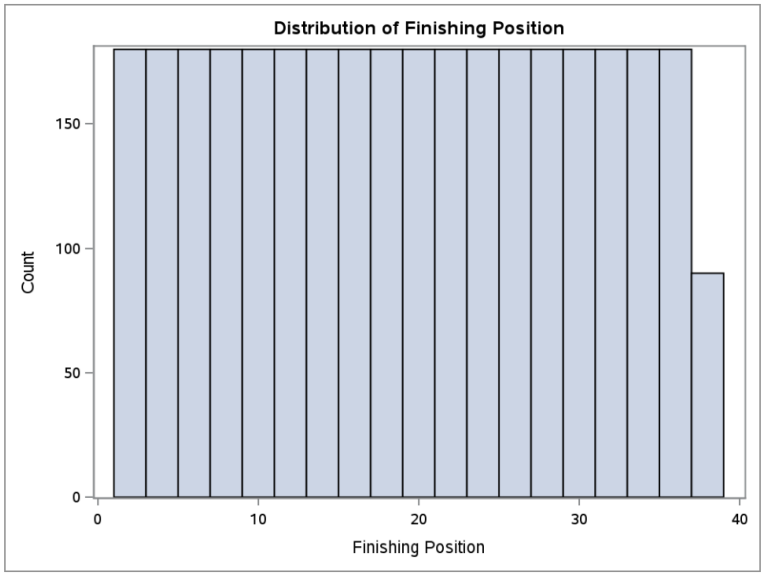

Figure 12: *finish* Histogram

The *finish* variable represents the driver's finishing position in the race. The boxplot and histogram indicate that this variable is unimodal, symmetric, and normally distributed. This should not be of much surprise, as each finishing position can only be unique to one driver.

points

Description: The total number of points that the driver earned throughout the race
Data Type: numeric
Minimum: 0
1st Quartile: 9
Median: 20
Mean: 20.86
3rd Quartile: 34
Maximum: 56

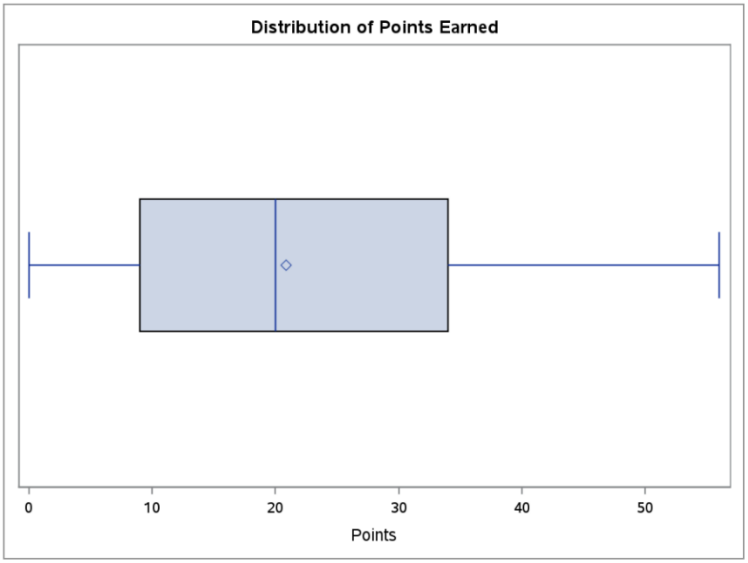

Figure 13: *points* Boxplot

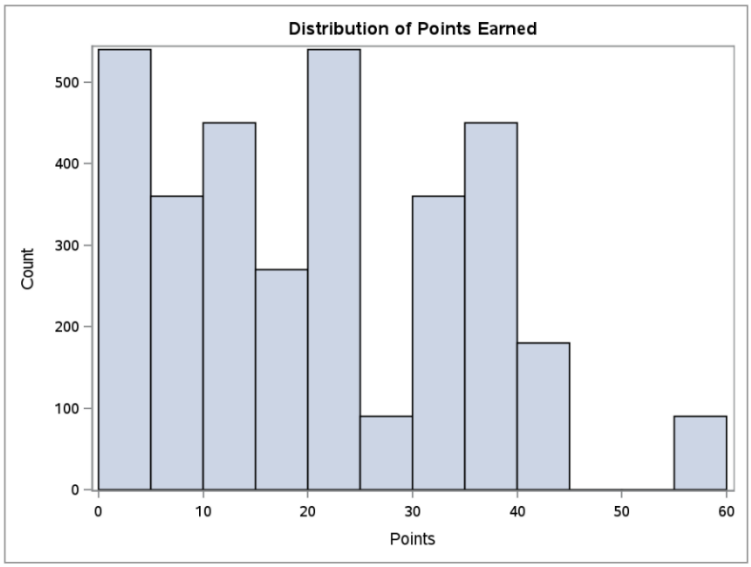

Figure 14: *points* Histogram

The *points* variable represents the total number of points that the driver earned during the race. Each race is broken into three stages: Stage 1, Stage 2, and the Final Stage. Drivers are eligible to receive points for their performances in Stages 1 and 2. Drivers running 1st through 10th at the conclusion of Stages 1 and 2 will receive points, starting with 10 points for 1st place, 9 points for 2nd place, down to 1 point for 10th place. Following the Final Stage, the driver in 1st (the race winner) receives 40 points, 2nd place 35, 3rd place 34, 4th place 33, and so on on a 35-to-2 scale. Those finishing 36th or greater are awarded 1 point. Points earned in Stages 1 and 2 are then added to what drivers earn after the Final Stage, summing to the total number of points earned during the race.

Looking at the boxplot and histogram for *points*, this variable appears to be positively-skewed, as the mean of 20.86 points is slightly greater than the median of 20 points. This distribution can likely be explained by the 5-point differential that 1st place has over 2nd place in Final Stage scoring, compared to just a 1-point advantage between all other neighboring finishing positions.

lap

Description: The lap number in the race
Data Type: numeric
Minimum: 1
1st Quartile: 23
Median: 45.5
Mean: 45.5
3rd Quartile: 68
Maximum: 90

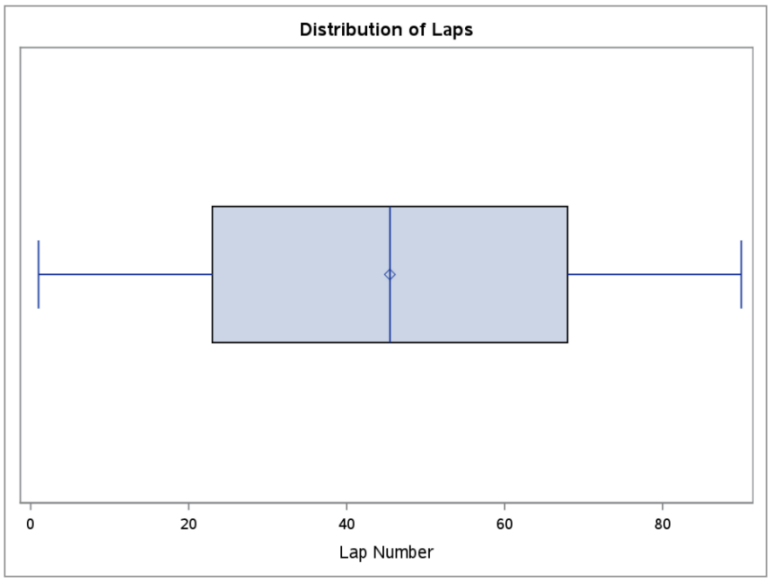

Figure 15: *lap* Boxplot

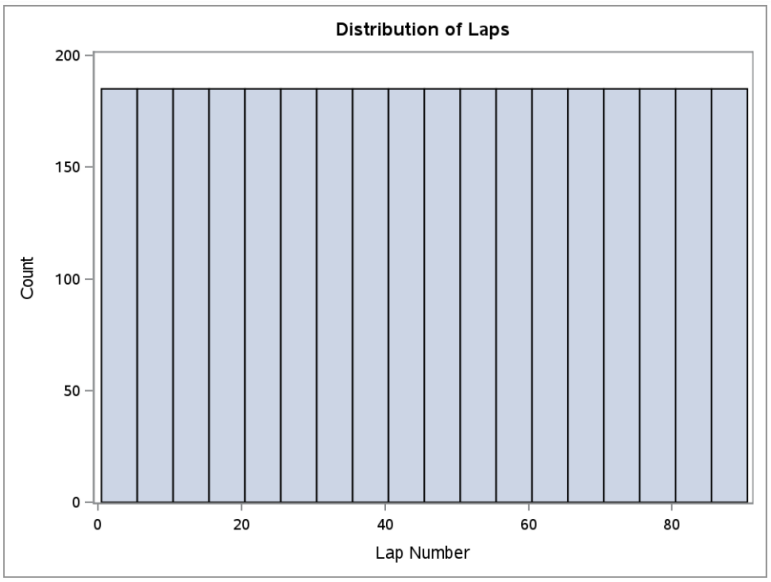

Figure 16: *lap* Histogram

The *lap* variable represents the lap number in the race. The boxplot and histogram indicate that this variable is unimodal, symmetric, and normally distributed. Such a distribution is to be expected, as each lap had a total of 37 observations for each driver that competed.

lap_time

Description: The overall time of the lap (in seconds)
Data Type: numeric
Minimum: 72.63
1st Quartile: 75.08
Median: 75.83
Mean: 88.93
3rd Quartile: 77.78
Maximum: 1,456.19

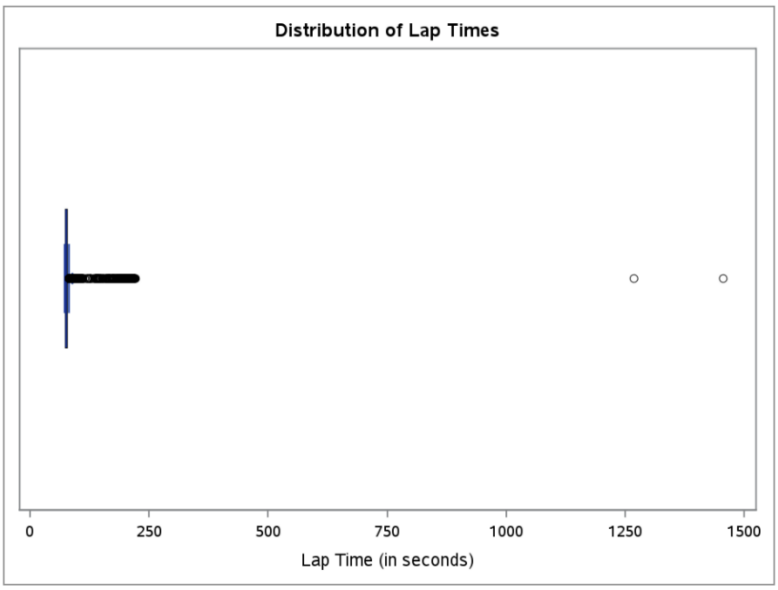

Figure 17: *lap_time* Boxplot

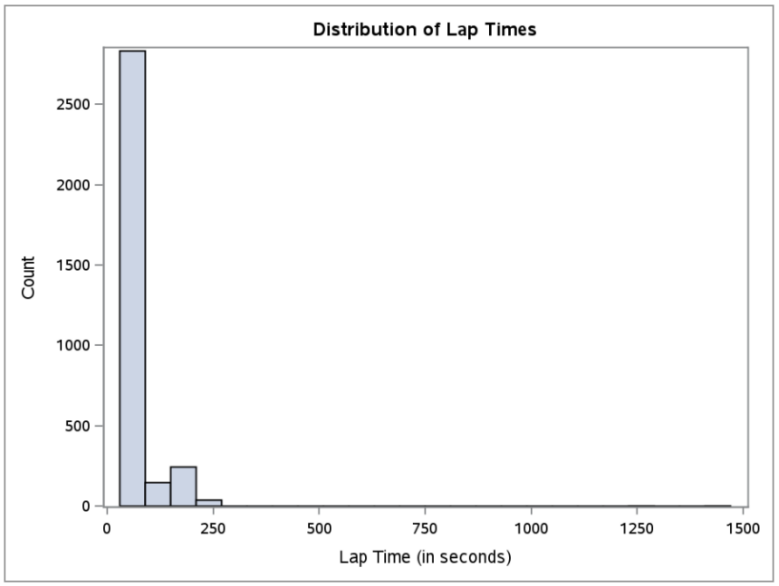

Figure 18: *lap_time* Histogram

The *lap_time* variable represents the overall time (in seconds) that it took the driver to complete a lap. Looking at the boxplot and histogram, this variable appears to be extremely positively skewed, as the mean of 88.93 seconds is significantly greater than the median of 75.83 seconds. This extreme skew can be explained by the cautions that occurred during the race. This race had a total of 4 cautions that lasted 6 laps. When a race is under caution, all cars are required to slow down and follow a pace car. The slow down causes their lap times to take much longer than if the race was under a green flag condition.

lap_speed

Description: The overall speed of the lap (in miles per hour)
Data Type: numeric
Minimum: 6.06
1st Quartile: 113.40
Median: 116.31
Mean: 108.08
3rd Quartile: 117.48
Maximum: 121.43

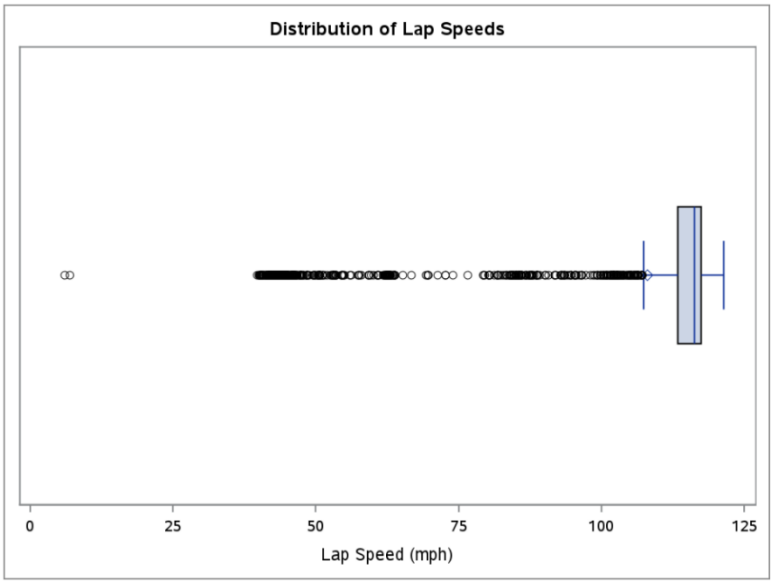

Figure 19: *lap_speed* Boxplot

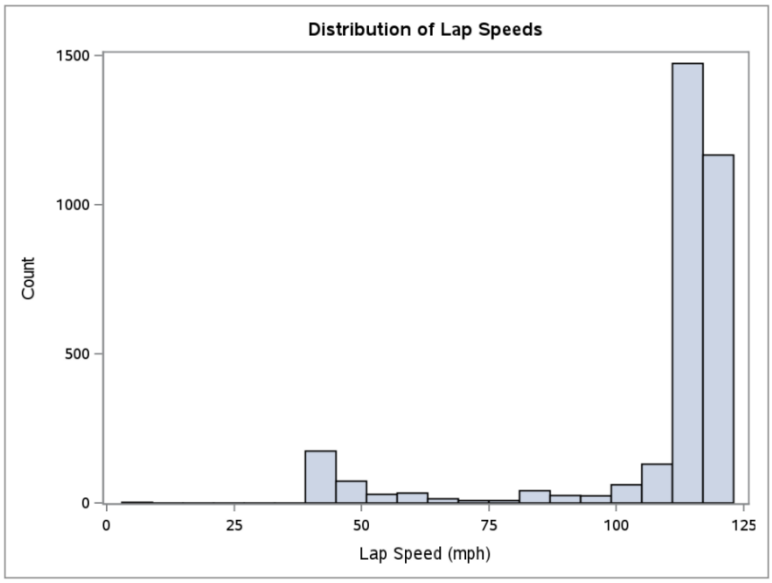

Figure 20: *lap_speed* Histogram

The *lap_speed* variable represents the overall speed (in miles per hour) that it took the driver to complete a lap. Looking at the boxplot and histogram, this variable appears to be negatively skewed, as the mean of 108.08 miles per hour is significantly less than the median of 116.31 miles per hour. As with *lap_time*, this skew can also be explained by the cautions that occurred during the race. This race had a total of 4 cautions that lasted 6 laps. When a race is under caution, all cars are required to slow down and follow a pace car. The slow down causes their lap speeds to be much lower than if the race was under a green flag condition.

running_position

Description: The position the driver was scored in for that lap
Data Type: numeric
Minimum: 1
1st Quartile: 10
Median: 19
Mean: 19
3rd Quartile: 28
Maximum: 37

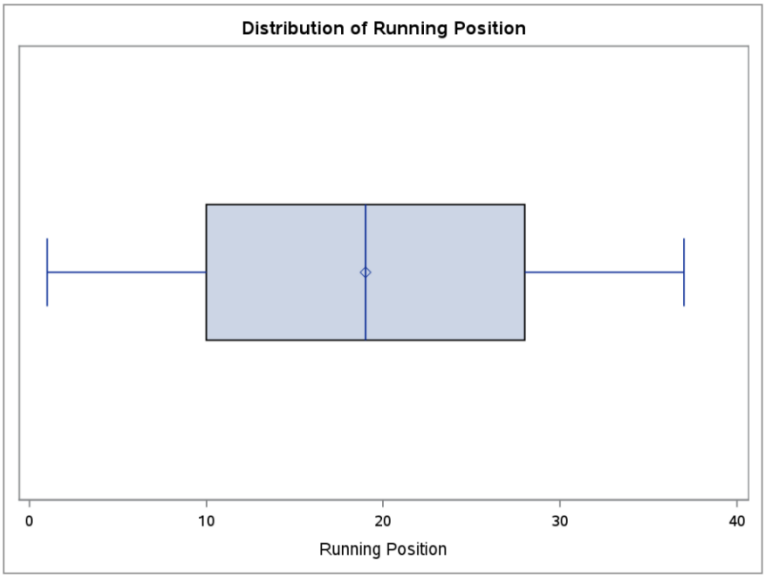

Figure 21: *running_position* Boxplot

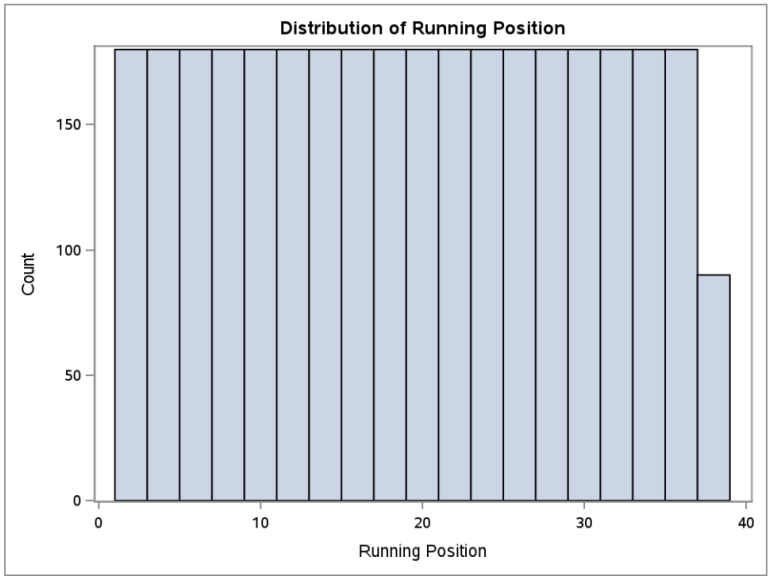

Figure 22: *running_position* Histogram

The *running_position* variable represents the position in which a driver was scored on a given lap. The boxplot and histogram indicate that this variable is unimodal, symmetric, and normally distributed. Such a distribution is to be expected, as each running position per lap can only be unique to one driver.

VARIABLE ANALYSIS

Start vs. Finish

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1748.85159	1748.85159	24.79	<.0001
Error	35	2469.14841	70.54710		
Corrected Total	36	4218.00000			

Root MSE	8.39923	R-Square	0.4146
Dependent Mean	19.00000	Adj R-Sq	0.3979
Coeff Var	44.20648		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.76577	2.81860	2.40	0.0218
Start	1	0.64391	0.12933	4.98	<.0001

Figure 23: Linear Model and Hypothesis Test

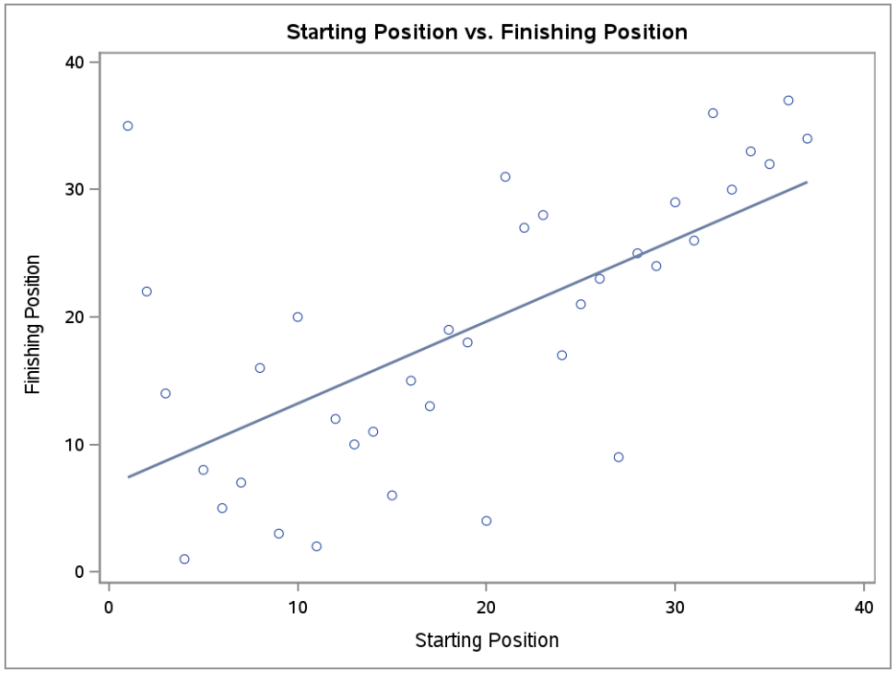

Figure 24: Scatter Plot

A simple linear regression was created using start as the lone predictor of finish. This yielded a regression equation of:

$$Finish = 6.76577 + 0.64391 \times Start$$

6.76577 is the y-intercept (β_0). It represents a driver's finishing position when *Start* (β_1) is equal to zero. The coefficient of the predictor, *Start*, can be interpreted as a 0.64391 increase in finishing position for each additional positional increase in starting position.

The null hypothesis was $H_0 : \beta_{Start} = 0$, the alternative hypothesis was $H_A : \beta_{Start} \neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 24.79, and the p-value was less than 0.0001. Based on the results of the test, we reject the null hypothesis because the p-value of < 0.0001 is less than the alpha of 0.05. There is statistically significant evidence that suggests β_{Start} is different from 0.

Start and *Finish* have a moderate, positive correlation coefficient of 0.6439. The coefficient of determination between these two variables is 0.4146, which means that approximately 41.46% of the variation in finishing position occurs because of changes in starting position. The weak predictive capability of *Start* indicates there are many more variables which can explain the other 58.54% of variation in a driver's finishing position.

Start vs. Points

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3600.42888	3600.42888	30.34	<.0001
Error	35	4153.89545	118.68273		
Corrected Total	36	7754.32432			

Root MSE	10.89416	R-Square	0.4643
Dependent Mean	20.86486	Adj R-Sq	0.4490
Coeff Var	52.21294		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	38.41892	3.65584	10.51	<.0001
Start	1	-0.92390	0.16774	-5.51	<.0001

Figure 25: Linear Model and Hypothesis Test

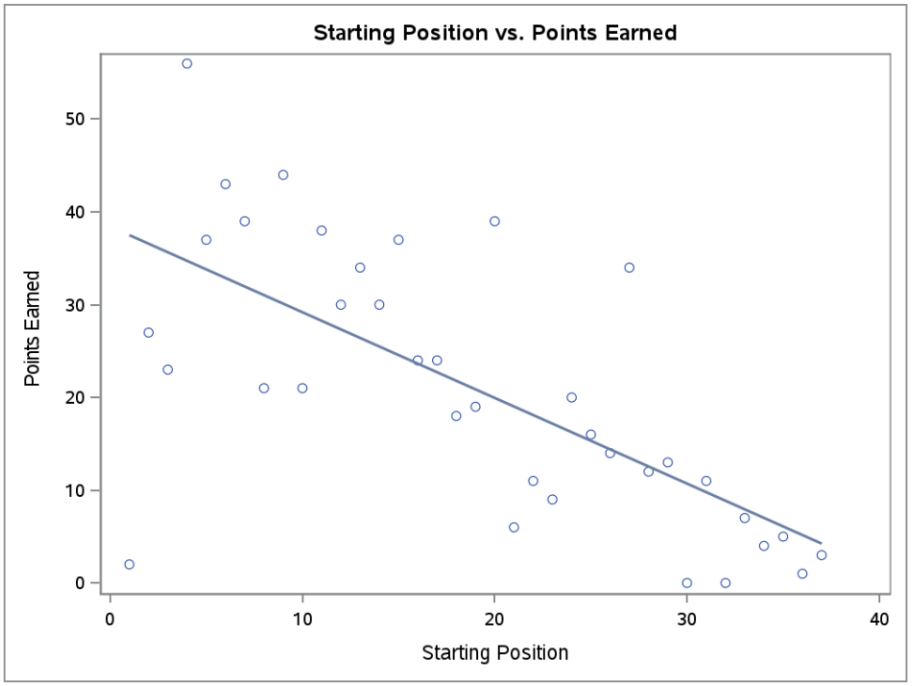

Figure 26: Scatter Plot

A simple linear regression was created using *Start* as the lone predictor of *Points*. This yielded a regression equation of:

$$Points = 38.41892 - 0.92390 \times Start$$

38.41892 is the y-intercept (β_0). It represents the points a driver earns when *Start* (β_1) is equal to zero. The coefficient of the predictor, *Start*, can be interpreted as a 0.92390 decrease in points earned for each additional positional increase in starting position.

The null hypothesis was $H_0 : \beta_{Start} = 0$, the alternative hypothesis was $H_A : \beta_{Start} \neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 30.34, and the p-value was less than 0.0001. Based on the results of the test, we reject the null hypothesis because the p-value of < 0.0001 is less than the alpha of 0.05. There is statistically significant evidence that suggests β_{Start} is different from 0.

Start and *Points* have a moderate, negative correlation coefficient of -0.6814. The coefficient of determination between these two variables is 0.4643, which means that approximately 46.43% of the variation in points earned occurs because of changes in starting position. The weak predictive capability of start indicates there are many more variables which can explain the other 53.57% of variation in the number of points that a driver earns.

Average Speed vs. Points

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5529.82615	5529.82615	87.01	<.0001
Error	35	2224.49817	63.55709		
Corrected Total	36	7754.32432			

Root MSE	7.97227	R-Square	0.7131
Dependent Mean	20.86486	Adj R-Sq	0.7049
Coeff Var	38.20907		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-802.86564	88.32004	-9.09	<.0001
Average_Speed	1	7.62818	0.81780	9.33	<.0001

Figure 27: Linear Model and Hypothesis Test

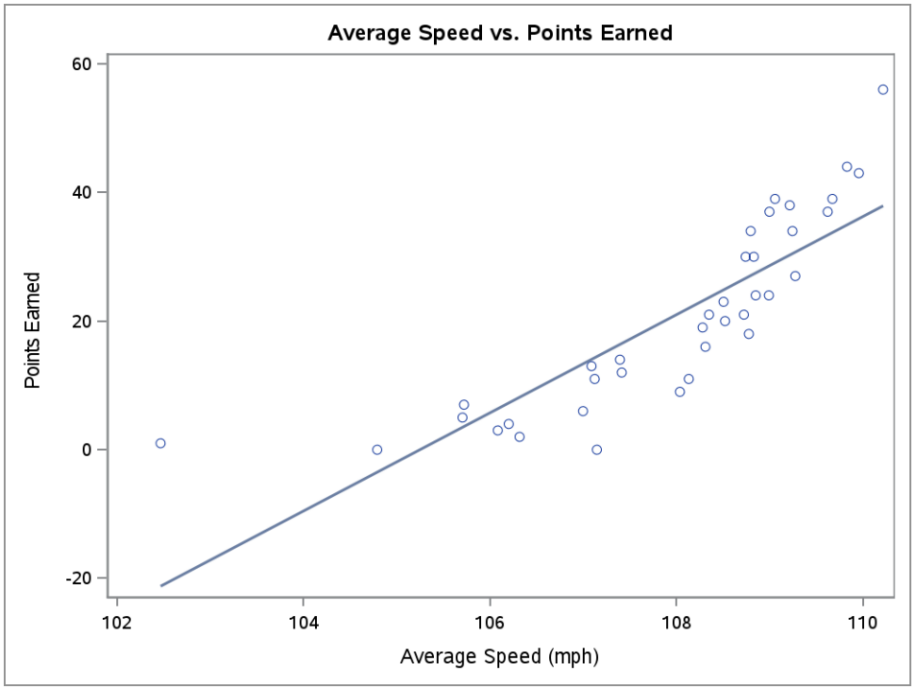

Figure 28: Scatter Plot

A simple linear regression was created using *Average_Speed* as the lone predictor of *Points*. This yielded a regression equation of:

$$Points = -802.86564 + 7.62818 \times Average\_Speed$$

-802.86564 is the y-intercept (β_0). It represents the number of points a driver earns when *Average_Speed* (β_1) is equal to zero. The coefficient of the predictor, *Average_Speed*, can be interpreted as a 7.62818 increase in points earned for each additional mile per hour increase in speed.

The null hypothesis was $H_0 : \beta_{Average\_Speed} = 0$, the alternative hypothesis was $H_A : \beta_{Average\_Speed} \neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 87.01, and the p-value was less than 0.0001. Based on the results of the test, we reject the null hypothesis because the p-value of < 0.0001 is less than the alpha of 0.05. There is statistically significant evidence that suggests $\beta_{Average\_Speed}$ is different from 0.

Average_Speed and *Points* have an extremely strong, positive correlation coefficient of 0.8445. The coefficient of determination between these two variables is 0.7131, which means that approximately 71.31% of the variation in points earned occurs because of changes in average speed. The moderate predictive capability of *Average_Speed* indicates there are several more variables which can explain the other 28.69% of variation in the number of points that a driver earns.

Average Speed vs. Finish

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3217.95948	3217.95948	112.62	<.0001
Error	35	1000.04052	28.57259		
Corrected Total	36	4218.00000			

Root MSE	5.34533	R-Square	0.7629
Dependent Mean	19.00000	Adj R-Sq	0.7561
Coeff Var	28.13333		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	647.37616	59.21777	10.93	<.0001
Average_Speed	1	-5.81910	0.54833	-10.61	<.0001

Figure 29: Linear Model and Hypothesis Test

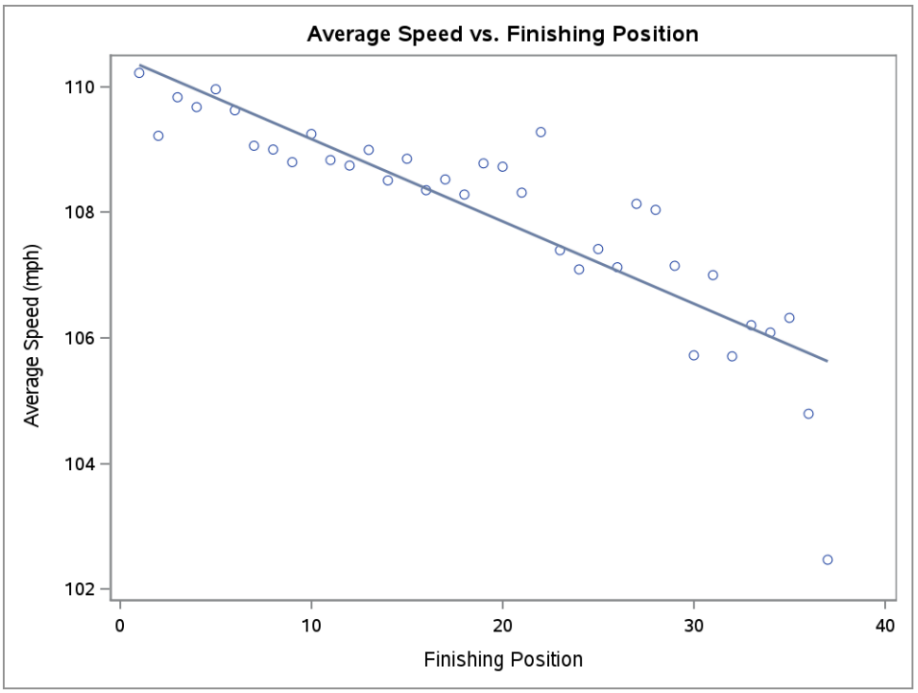

Figure 30: Scatter Plot

A simple linear regression was created using *Average_Speed* as the lone predictor of *Finish*. This yielded a regression equation of:

$$Finish = 647.37616 - 5.81910 \times Average\_Speed$$

647.37616 is the y-intercept (β_0). It represents a driver's finishing position when *Average_Speed* (β_1) is equal to zero. The coefficient of the predictor, *Average_Speed*, can be interpreted as a 5.81910 decrease in finishing position for each additional mile per hour increase in speed.

The null hypothesis was $H_0 : \beta_{Average\_Speed} = 0$, the alternative hypothesis was $H_A : \beta_{Average\_Speed} \neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 112.62, and the p-value was less than 0.0001. Based on the results of the test, we reject the null hypothesis because the p-value of < 0.0001 is less than the alpha of 0.05. There is statistically significant evidence that suggests $\beta_{Average\_Speed}$ is different from 0.

Average_Speed and *Finish* have an extremely strong, negative correlation coefficient of -0.8734. The coefficient of determination between these two variables is 0.7629, which means that approximately 76.29% of the variation in finishing position occurs because of changes in average speed. The moderate predictive capability of *Average_Speed* indicates there are several more variables which can explain the other 23.71% of variation in a driver's finishing position.

Average Running Position vs. Finish

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3748.73880	3748.73880	279.60	<.0001
Error	35	469.26120	13.40746		
Corrected Total	36	4218.00000			

Root MSE	3.66162	R-Square	0.8887
Dependent Mean	19.00000	Adj R-Sq	0.8856
Coeff Var	19.27169		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.55638	1.47718	-2.41	0.0215
Average_Running_Position	1	1.18718	0.07100	16.72	<.0001

Figure 31: Linear Model and Hypothesis Test

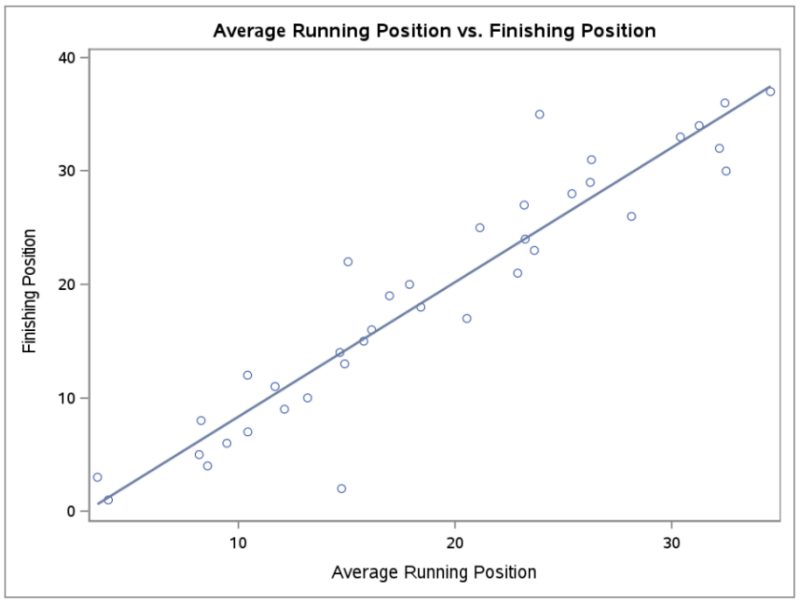

Figure 32: Scatter Plot

The average running position is an interesting key performance indicator (KPI) in that it is reflective of a driver's entire race, not just where they finished. It is calculated for each driver by summing their running positions for each lap and dividing them by the total number of laps.

A simple linear regression was created using *Average_Running_Position* as the lone predictor of *Finish*. This yielded a regression equation of:

$$Finish = -3.55638 + 1.18718 \times Average\_Running\_Position$$

-3.55638 is the y-intercept (β_0). It represents a driver's finishing position when *Average_Running_Position* (β_1) is equal to zero. The coefficient of the predictor, *Average_Running_Position*, can be interpreted as a 1.18718 increase in finishing position for each additional positional increase in average running position.

The null hypothesis was $H_0 : \beta_{Average\_Running\_Position} = 0$, the alternative hypothesis was $H_A : \beta_{Average\_Running\_Position} \neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 279.60, and the p-value was less than 0.0001. Based on the results of the test, we reject the null hypothesis because the p-value of < 0.0001 is less than the alpha of 0.05. There is statistically significant evidence that suggests $\beta_{Average\_Running\_Position}$ is different from 0.

Average_Running_Position and *Finish* have an extremely strong, positive correlation coefficient of 0.9427. The coefficient of determination between these two variables is 0.8887, which means that approximately 88.87% of the variation in finishing position occurs because of changes in average running position. The high predictive capability of *Average_Running_Position* indicates there are likely only a few more variables which can explain the other 11.13% of variation in a driver's finishing position.

Average Running Position vs. Points

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6914.12262	6914.12262	288.02	<.0001
Error	35	840.20170	24.00576		
Corrected Total	36	7754.32432			

Root MSE	4.89957	R-Square	0.8916
Dependent Mean	20.86486	Adj R-Sq	0.8886
Coeff Var	23.48238		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51.49828	1.97660	26.05	<.0001
Average_Running_Position	1	-1.61228	0.09500	-16.97	<.0001

Figure 33: Linear Model and Hypothesis Test

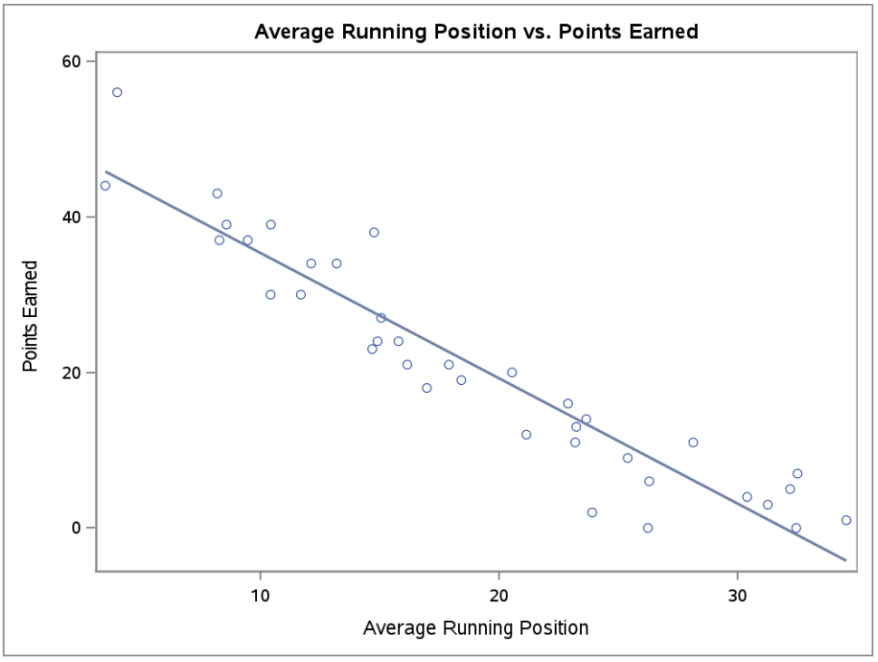

Figure 34: Scatter Plot

A simple linear regression was created using *Average_Running_Position* as the lone predictor of *Points*. This yielded a regression equation of:

$$Points = 51.49828 - 1.61228 \times Average\_Running\_Position$$

51.49828 is the y-intercept (β_0). It represents the number of points a driver earns when *Average_Running_Position* (β_1) is equal to zero. The coefficient of the predictor, *Average_Running_Position*, can be interpreted as a 1.61228 decrease in points earned for each additional positional increase in average running position.

The null hypothesis was $H_0 : \beta_{Average\_Running\_Position} = 0$, the alternative hypothesis was $H_A : \beta_{Average\_Running\_Position} \neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 288.02, and the p-value was less than 0.0001. Based on the results of the test, we reject the null hypothesis because the p-value of < 0.0001 is less than the alpha of 0.05. There is statistically significant evidence that suggests $\beta_{Average\_Running\_Position}$ is different from 0.

Average_Running_Position and *Points* have an extremely strong, negative correlation coefficient of -0.9442. The coefficient of determination between these two variables is 0.8916, which means that approximately 89.16% of the variation in points earned occurs because of changes in average running position. The high predictive capability of *Average_Running_Position* indicates there are likely only a few more variables which can explain the other 10.84% of variation in the number of points that a driver earns.

CONCLUSION

SUMMARY

The overarching question for this analysis was "Which factors best explain driver success on the race track?", with success being defined in terms of finishing position and points earned. After some investigation, it is evident that the top factors explaining success are average speed and average running position. Average speed had the second best correlation coefficients for any variable, with values of -0.87 and 0.84 for finishing position and points earned, respectively. Average speed also had the second best coefficients of determination out of any variables, with values of 0.76 and 0.71 for finishing position and points earned, respectively. Average running position had the best correlation coefficients for any variable, with values of 0.94 and -0.94 for finishing position and points earned, respectively. Average running position also had the best coefficients of determination out of any variables, with values of 0.89 and 0.89 for finishing position and points earned, respectively.

Although these two variables had extremely high correlations and R-squares, the results of the inferential analysis and hypothesis tests were a bit alarming. In every hypothesis test, the null hypothesis was rejected, meaning there was statistically significant evidence that suggested β_1 was different from 0. This was true whether β_1 was the starting position, average speed, or average running position. The implications of these results indicated that no lone predictor was useful in predicting the response. Therefore, a simple linear regression was not an appropriate model to use in this data set.

LIMITATIONS

The first limitation of this analysis was that it was conducted on a relatively small sample size. The data in this analysis was representative of only one of the thirty-six races that occurred during the 2021 NASCAR Cup Series season. At ninety laps total, this race also had the fewest number of scheduled laps for any race during the season. Having a smaller sample size made the central limit theorem less applicable to the data, as it was not as reasonable to assume normality or equal variance in the data. Something else worth noting is that this specific race took place on a road course. Therefore, the findings of this study likely aren't useful in assessing driver success at other types of tracks, such as short tracks, intermediate tracks, and superspeedways. The final limitation of this analysis is the season and time of day in which the race occurred. This race occurred on a sunny summer day, which means that the track was hotter and slicker than usual, making for less grip, worse handling, and a looser driving condition in the car. It would not be ideal to use this data to predict results in a race occurring during a different season and/or time of day.

NEXT STEPS

This analysis could be improved by conducting a multiple linear regression to determine whether more than one variable is better at predicting driver success. It would be interesting to explore whether combining variables like average speed and average running position would improve the correlations and coefficients of determination for finishing position and points earned. It would also be interesting to conduct hypothesis testing on these multiple linear regressions to determine if the null hypothesis instead failed to be rejected. Such results could indicate that a multiple linear regression would be a more appropriate model to use on this data set. In conducting a multiple linear regression, it would be important to also conduct tests for multicollinearity and variance inflation factors in the predictor variables. Finally, this analysis could also be improved by applying the models to different races altogether to determine whether the same findings are prevalent. Making an adjustment like this would prove the analysis to be more scalable and robust.

ABOUT THE AUTHOR

Blake Pappas is a Lead Operations Analyst at Techtronic Industries (TTI). He graduated from Anderson University in 2021 with a Bachelor of Science in Financial Analysis. Blake currently attends Clemson University, where he is a graduate student in their Data Science and Analytics program. Born and raised in Egg Harbor Township, NJ, Blake is the youngest of five children. His favorite hobbies consist of running, reading, and sports. He has hopes to one day start his own consulting company that specializes in data analytics and information engineering for small businesses. He lives in Mauldin, SC.

REFERENCES

1. “2021 Go Bowling at The Glen.” Racing Reference, https://www.racing-reference.info/race-results/2021_Go_Bowling_at_the_Glen/W/. Accessed 13 October 2022.

2. “2021 NASCAR Cup Series Results.” Jayski, https://www.jayski.com/nascar-cup-series/2021-nascar-cup-series-results/. Accessed 13 October 2022.

3. “Go Bowling at The Glen: Race Recap.” NASCAR, https://www.nascar.com/results/racecenter/2021/nascar-cup-series/go-bowling-at-the-glen/stn/recap/. Accessed 12 October 2022.

4. Pfitzner, Barry C. & Rishel, Tracy D. “Do Reliable Predictors Exist for the Outcomes of NASCAR Races?” The Sport Journal, https://thesportjournal.org/article/do-reliable-predictors-exist-for-the-outcomes-of-nascar-races/. Accessed 13 October 2022.

5. “Watkins Glen International: Home.” Watkins Glen International, https://www.theglen.com/. Accessed 7 October 2022.

6. "Watkins Glen International Road Course." syracuse.com, https://www.syracuse.com/resizer/7S15ydIrlyYui3a5L2adL0L4X4k=/800x0/smart/cloudfront-us-east-1.images.arcpublishing.com/advancelocal/2ZJLQZ7J4ZGNXF3XKLWDAMPNPQ.jpeg. Accessed 15 October 2022.