CURSO 2019-2020

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

Memoria final

Carreras de Montaña

ETS de Ingeniería Informática □ Camino de Vera, s/n 46022 Valencia
https://www.upv.es/titulaciones/GCD

ÍNDICE

Autores

Nombres y apellidos de los autores:

Antonio	Cortés Sanz
Oscar	Sebastián Mollá
Cándido	García Rodríguez
Pablo	Atienza Valero
Jorge David	Mínguez Fons

1. Alcance del proyecto

Nuestro estudio trata de determinar qué factores afectan al rendimiento deportivo en las carreras de montaña de larga distancia, así como intentar atraer a estas carreras un determinado tipo de corredores. analizaremos para ello ciertas variables (temperatura, género, edad, precipitaciones, año de realización de la carrera ...) para observar sus efectos. Analizar estos agentes por separado y entre sí y explicar los resultados. Además estudiar los patrones de aquellos corredores que mejor lo han hecho nos puede dar una visión de la correcta interpretación de dichos factores y cómo sobrellevarlos.

REQUISITOS:

- Expresar claramente el conjunto de tareas que vamos a realizar con el fin de alcanzar los objetivos del proyecto.
- Tratar de conjunto de datos para obtener la información más detallada y precisa posible.
- Nuestra materialización de los datos ha de ser realista y coherente con los objetivos y nuestros recursos.

RESTRICCIONES:

- Desconocimiento al principio del proyecto sobre programación en R.
- Datos ausentes en la base de datos, no tan solo de gente que se deja la carrera si nos algunas mediciones que no han sido tomadas.
- Datos anómalos de tiempos, que a la hora de agregar datos provocaba que salieran errores.
- No se ha podido averiguar cómo aplicar web scraping en páginas web dinámicas usando R o python.

Asunciones y supuestos al inicio del proyecto

En primer lugar queremos aclarar que tenemos unos conocimientos 'limitados', por tanto no tenemos la habilidad ni la experiencia para desarrollar un trabajo más profesional y preciso, dicho esto pasamos a las asunciones.

Asunciones:

- Todos los miembros del equipo se comprometen a buscar información e indagar sobre el tema principal de nuestro proyecto, así como contrastarlo con los resultados obtenidos.
- Suponemos que la temperatura, y las precipitaciones son factores que van a afectar a los corredores, y a sus rendimiento en las carreras
- Asumimos que los hombres de media son más rápidos que las mujeres.
- Suponemos que los participantes más jóvenes son más veloces que los participantes con edades más avanzadas.
- Se cree que habrá una población de hombres y mujeres similar en las 3 carreras estudiadas.
- La importancia de aprender y aplicar nuevas técnicas que puedan ayudar a precisar los resultados del proyecto tanto estadísticas como de programación, web scraping…

Productos entregables

En lo que al producto final se refiere, nuestro equipo pretende entregar un trabajo bien realizado y coherente, queremos que este sea un trabajo completo, teniendo en cuenta nuestra falta de experiencia y de conocimientos sobre el análisis de datos, otros de los productos entregables de gran importancia que hemos realizado es la entrega de HITOS, estos pueden ser utilizados para apreciar la evolución que ha sufrido nuestro trabajo.

Límites del proyecto (qué haremos y qué no haremos)

No podremos estudiar la influencia de determinados factores como el viento, ya que no podíamos saber muy bien la orientación del corredor, ni aislar su efecto, ya que otros factores entran en juego, como el desnivel y el efecto escudo que ocurre debido a las montañas. Tampoco hemos podido analizar si el bajo rendimiento se debe a razones externas como lesiones, falta de equipo, ni tampoco saber a qué nivel compiten los corredores y si se dedican profesionalmente a ello o simplemente es un hobby, podemos suponer que los mejores corredores se dedican profesionalmente a las carreras pero no lo podemos asegurar ya que existen aficionados con un nivel muy elevado.

Resultados esperados

Esperamos obtener resultados concluyentes en cuanto el clima, ya que son carreras de larga duración, y este factor puede llegar a afectar mucho a los corredores, además del factor geográfico, ya que la distribución de la elevación a lo largo del circuito, y como el corredor raciona sus fuerzas puede ser un factor fundamental en sus resultados.

Criterios de éxito del proyecto y criterios de aceptación del producto

Todos los estudios se emprenden con ilusión y con el objetivo de conseguir cubrir una necesidad, desafortunadamente no siempre se consigue. De hecho, en muchas ocasiones todo el esfuerzo que se dedica por parte del equipo no fructifica en la satisfacción de los requisitos del 'cliente'. Es por esto conocer qué criterios pueden llevar al éxito a nuestro proyecto es de vital importancia .

Deberíamos tener en cuenta aspectos como:

- **Eficiencia**: en lo relativo al cumplimiento de los plazos, coste de oportunidad, uso de recursos.
- **Impacto en el cliente:** si el producto del proyecto cumple unas necesidades reales.
- **Impacto en el equipo**: conseguir que el equipo crezca, se sienta involucrado y valorado y aumente sus competencias es la mejor forma de mantener el talento en la organización.
- **Preparación para el futuro:** lo que se aprende en cada proyecto sirve para hacer más eficiente la ejecución del siguiente proyecto.

1.1. Objetivos del proyecto

Desde el inicio del trabajo nuestro objetivo ha sido estudiar los factores que afectan al rendimiento deportivo en carreras de montaña con el fin de ayudar a aquellos deportistas, tanto profesionales como aficionados, que no están obteniendo los resultados deseados o se han estancado, a la par que ayudar a los organizadores a atraer un determinado público a las dichas carreras. Al mismo tiempo pensamos que realizar dicho estudio podría ser útil ya que en los últimos años el deporte ha pegado un paso de gigante y está a la orden del día en la vida de mucha y cada vez más personas, el objetivo de todo esto también es para motivar a la gente a seguir con el deporte ya que es muy importante para disfrutar, para mejorar la salud y para un montón de cosas más. Según la UE, España es uno de los países con mayor afición por este deporte. Con el fin de estructurar nuestro proyecto hemos dividido nuestros objetivos en: generales y específicos.

En los últimos años, en nuestro país, como ya hemos dicho, ha habido un fuerte crecimiento por el amor a este deporte, una de las más importantes y que sigue activa últimamente es como mejorar en este tipo de deporte, con todos sus factores y variables diferentes. Por este motivo hemos decidido establecer como objetivo general "¿Qué factores influyen en el rendimiento del corredor?".

Para alcanzar el objetivo general, vamos a estudiar por separado estos objetivos.

- Analizar qué etapas son más cruciales durante la carrera para obtener mejores resultados.

- Predecir si un corredor va a abandonar o proseguir su carrera dependiendo de su actuación en ciertas etapas

- Estudiar la internacionalidad de las carreras, si una carrera tiene presencia internacional será debido en gran medida a que es una buena carrera.

- Determinar si un corredor está preparado para participar en según qué carreras y si no lo está , que debe hacer para estarlo
- Estudiar e intentar explicar las velocidades medias y los tiempos de espera de las etapas en función de la orografía de la etapa.
- Estudiar cómo puede llegar a afectar la climatología en las velocidades y tiempos de espera.
- Ver si factores como la edad o el género influyen en los resultados.
- Analizar los mejores, peores y tiempos medios a lo largo de las últimas décadas

1.2. Limpieza de datos

La limpieza está explicada más minuciosamente en el apartado del Hito 2 sobre limpieza que se puede encontrar aquí
https://drive.google.com/file/d/1ZndEEtaLrVwNo6VwR6atZFQMLtdBggGB/view?usp=sharing

COUNTER CLOCKWISE

Tras eliminar los datos faltantes de edad sexo y procedencia que representaban un 7% del total, hemos pasado a eliminar los valores faltantes en los tiempos registrados durante la carrera, con lo que eliminamos aproximadamente un 16% más de los datos, al finalizar la limpieza contamos con 1151 corredores.

CLOCKWISE

Para esta base de datos hemos detectado 116 valores faltantes para la variable edad, hemos eliminado las filas que contienen estos valores faltantes, ya que solo representan un 8% de nuestros datos y creemos que no va a repercutir en el análisis, además al estudiar valores faltantes de los tiempos recopilados, observamos un 11,5 % de filas con datos faltantes, por lo que las eliminamos y finalizamos con una base de 1339 corredores.

BEAR

No hay datos faltantes en cuanto a edad y sexo en esta base de datos, sin embargo hay muchos datos faltantes en los tiempos recogidos, casi un 30% de la base de datos , que eliminamos con lo que pasamos a tener 1538 corredores.

Además en esta carreras se tuvieron que eliminar 6 corredores debido a que salían velocidades negativas, es decir había un error de transcripción por parte de la página web de donde se obtuvieron los datos. Por lo cual se quedaron un total de 1532 corredores.

1.3. Utilidad del estudio

Además de ser un estudio académico que pueda llegar a servir a gente/grupos de estudio que se propongan objetivos parecidos con bases de datos de competiciones con un cierto desgaste físico, puede servir a atletas que participan en estos por tal de poder llegar a hacerse una idea de a qué nivel deben de rendir para poder llegar a la meta deseada en estas carreras.

Sin olvidar que es un estudio didáctico de la asignatura de Proyecto 2 por lo que la principal utilidad es que los integrantes aprendamos cómo trabajar en equipo para sacar un proyecto adelante.

2. Configuración del proyecto

2.1. Fuentes de datos

Los datos que finalmente se han obtenido proceden de la página web OpenSplitTime, que es una página web especializada en carreras de montaña. Respectivamente las páginas web son:

- The Bear 100 Course
- Hardrock 100 Counter-Clockwise
- Hardrock 100 Clockwise

Integración y transformación de los datos

https://drive.google.com/file/d/1ZndEEtaLrVwNo6VwR6atZFQMLtdBggGB/view?usp=sharing

Hemos actualizado los apartados del hito 2 sobre integración, como se puede ver ahí, no hemos integrado las bases en una sola, ya que vamos a estudiar las carreras por separado y luego analizar los resultados, tampoco integramos los datos del clima. En los apartados de Transformación, mostramos como hemos modelizado y calculado los datos agregados(velocidad media, tiempo de descanso) para las carreras, y también para los datos del clima(temperatura relativa).

Herramientas y técnicas utilizadas

A lo largo del proyecto se han utilizado muchas técnicas y herramientas de trabajo, pero los podemos dividir en 2 grandes grupos que coinciden a su vez con los 2 lenguajes de programación que se han tenido que usar para el desarrollo de este proyecto:

- **R:** Hemos usado librerías como dplyr y tidyverse para limpiar y modelizar nuestros datos, como no disponíamos de datos iniciales, hemos tenido que aprender cómo extraer los datos de eventos de OpenSplitTime utilizando R, que nos ha sido muy fructífero, y es un conocimiento muy útil para futuros proyectos. Además de toda la integración y el análisis de los datos se ha realizado en este programa.

- **Python:** hemos utilizado principalmente la librería pandas para tratar nuestros datos, también hemos usado las librerías gpx y minidom para parsear los archivos GPX (archivos de información de mapas de GPS) de los circuitos de la carrera y la librería matplotlib para representar la información resultante.

Los códigos que hemos utilizado se encuentran disponibles para su visualización en esta cuenta de Kaggle.

3. Análisis

3.1.. Analisis Clima

Para estudiar las condiciones climáticas durante la carrera se extrajeron datos referentes a las condiciones metereológicas de Weather Underground, una web que recopila datos climáticos en tiempo real e históricos, para diferentes estaciones repartidas por el mundo.

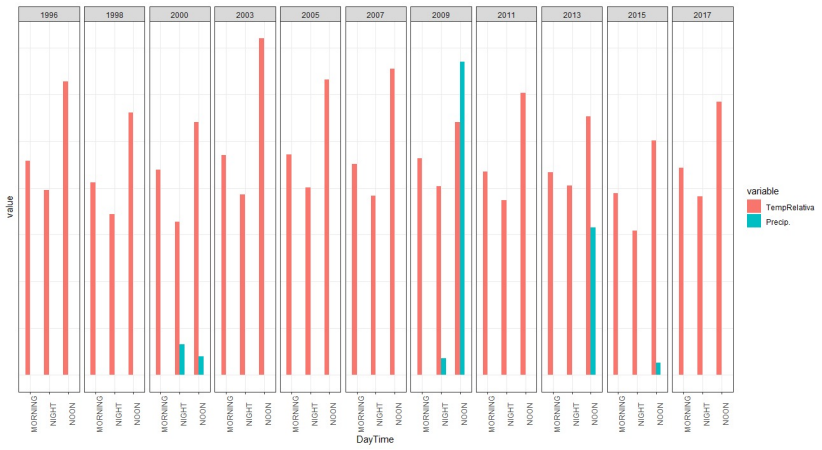
Inicialmente y tal como nos sugirieron, buscamos estaciones cercanas al terreno donde iban a acontecer los eventos, pero que además fuesen estaciones con localizaciones elevadas, ya que la carrera ocurre en ocasiones a grandes altitudes, sin embargo nos topamos con una limitación, ya que los puestos que cumplían las condiciones que buscábamos, no contaban con la mayoría de datos históricos, por lo que decidimos aproximar el clima según las mediciones tomadas por puestos cercanos al circuito que contasen con un registro aceptable, obviando su elevación.

Tras estudiar varias opciones nos dimos cuenta que la elección más acertada era elegir estaciones posicionadas en algún aeropuerto cercano, ya que debido a los vuelos, y la gran influencia del factor meteorológico en ellos, tenían registros completos de los años que deseábamos.

HardRock -> Durango-La Plata County Airport

Bear -> Logan-Cache-Station Smithfield UT

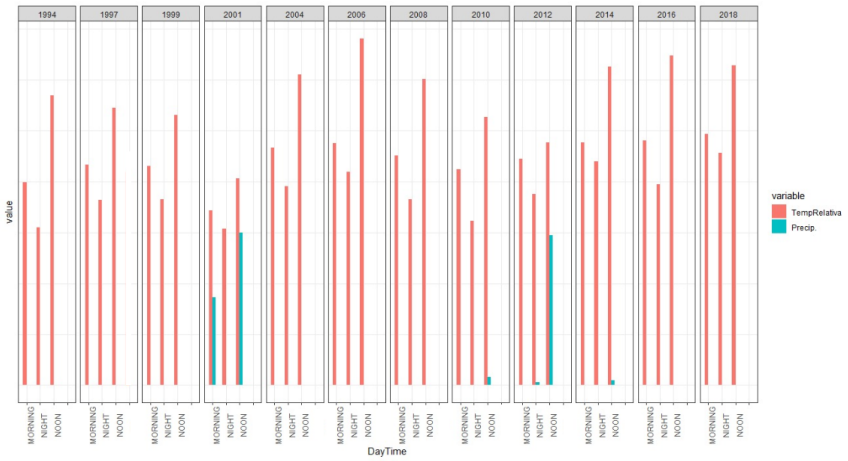
Hardrock counter-clockwise



Para la carrera Hardrock para los años en los cuales el circuito Counter Clockwise se utiliza, podemos ver temperaturas elevadas, sobretodo en las tardes del evento en 2003, el evento del 2015 es en el que las temperaturas son menores ,también en cuanto a las temperaturas vemos un patrón siendo las tardes más calurosas que las mañanas. y las mañanas evidentemente más calurosas que las noches.

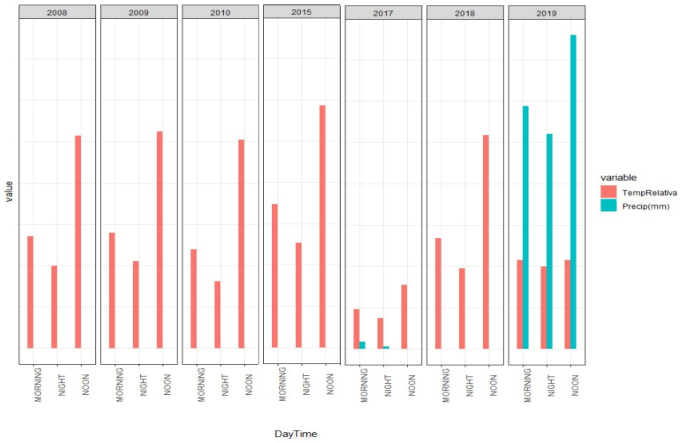
Al ser el mes de julio, es lógico observar que solo en cuatro de los once llueve algo, es más, en dos de ellos, estos solo presentan una precipitación mínima, en 2009 y 2013, podemos ver como en los atardeceres existen grandes precipitaciones.

Clockwise



Al igual que las ediciones de HardRock con el circuito Counter Clockwise, podemos ver temperaturas elevadas y pocas lluvias, solo en 4 ediciones llueve, y solo en dos de ellas la precipitación es notable.

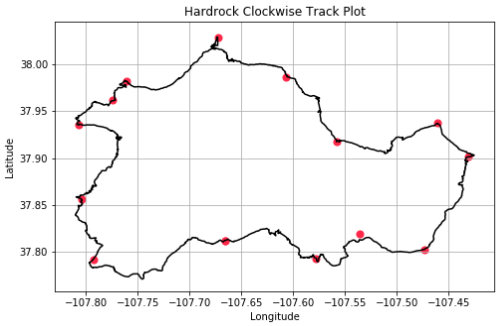
The Bear

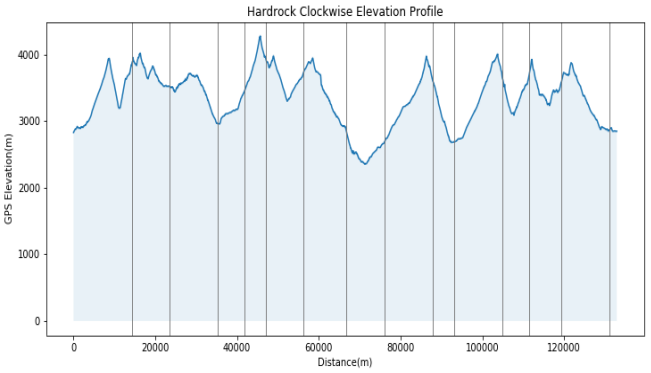


Esta carrera , al realizarse en septiembre , debido al mes en el cual ocurre tiene temperaturas mucho más bajas que las de Hard Rock, también observamos mucha más precipitaciones que las dos anteriores realizadas en junio,podemos ver que puede llegar a llover durante la mañana, tarde o noche indistintamente, vemos que para el año 2014 la lluvia fue bastante más homogénea que en 2016 donde podemos ver que hubo más precipitaciones durante las noche, para finalizar, el año con más precipitaciones fue el año 2019.

3.2. Análisis de las carreras

Clockwise

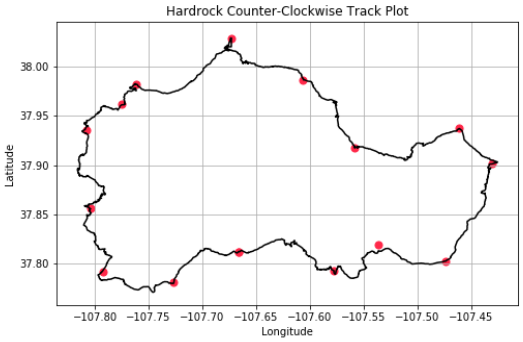


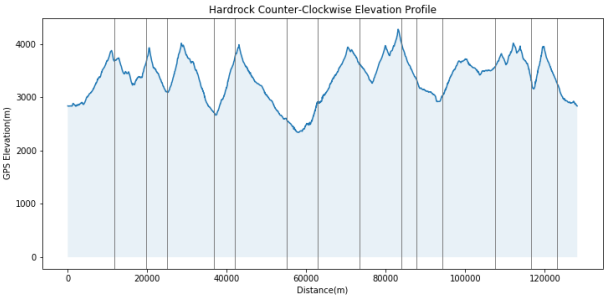


Se puede ver que la elevación inicial de esta carrera es de 3000 metros por encima del nivel del mar,llegando hasta un máximo de 5000 metros.

En el gráfico anterior se analiza la forma y coordenadas del recorrido del circuito para la carrera Clockwise(primer gráfico) con su respectiva elevación cada metro de la carrera, como podemos ver la carrera tiene muchos altibajos como era de esperar ya que se trata de carreras de montaña.

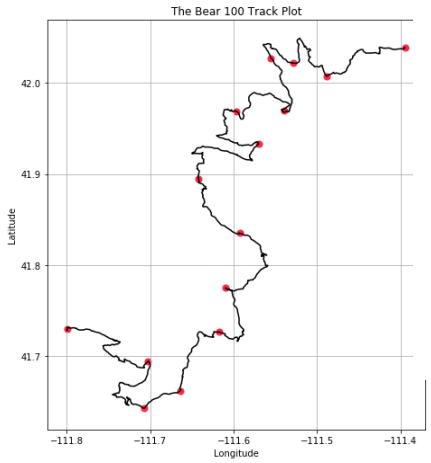
Counter clockwise

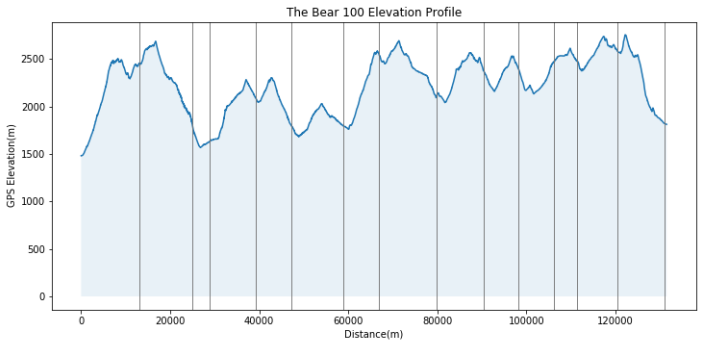




En la carrera counter-clockwise (gráfico de la izquierda) vemos que la carrera tiene un circuito bastante similar en cuanto a forma,con los mismos picos ,podemos concluir por tanto que la dificultad de una o la otra son prácticamente iguales por lo que los participantes que se les de bien una de estas , podrá participar sin problemas en la otra,cosa que les puede interesar ya que en este tipo de carreras son conocidas por otorgar premios en metálico de hasta 30000 dólares.

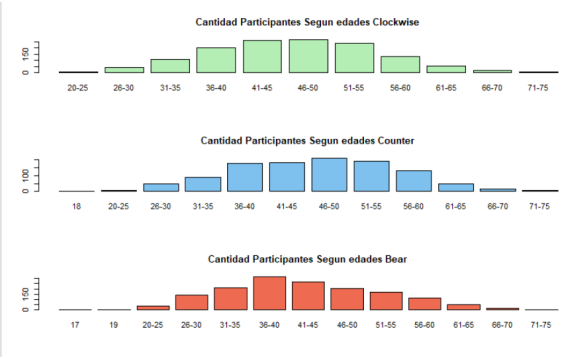
The Bear





Para la carrera de Bear vemos un importante cambio en lo que es la forma del circuito , siendo este una línea , también vemos que el desnivel es menos marcado en esta carrera que en las dos anteriores , aunque también es alto ya que se trata de una carrera de montaña , cabe destacar que la altitud es mucho menor siendo el pico más alto unos 3000 m mientras que en las anteriores está sobre los 4000 metros, como no podemos analizar la influencia de factores como el viento en estas carreras por falta de datos , concluimos que esta carrera es ,en principio, más fácil que las anteriores, al tener una elevación menor en toda la carrera y unos cambios menos bruscos, o lo que es lo mismo , hay menos subidas y bajadas y estas son menos bruscas que en las anteriores dos carreras.

En conclusión si observamos las distribuciones por edades en los tres circuitos podemos observar un claro patrón existe una pequeña cantidad de corredores en el primer cuartil con edades muy jóvenes o elevadas, la mayoría de estos corredores tienen unas edades que oscilan entre los 30 y los 50 años.



(Para ver las correlaciones sobre el género, edad, tiempo final, descanso , velocidad medio y año ir al apartado 1.5).

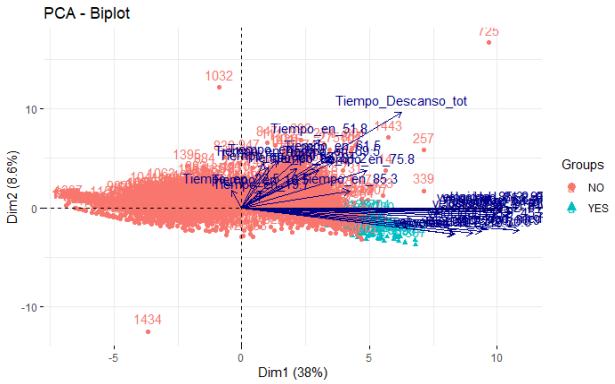
4. Resultados

¿Qué etapas son más cruciales durante la carrera para obtener mejores resultados, y qué diferencias existen en los corredores TOP y el resto?

Para estudiar el efecto de las variables en la explicación de la variabilidad de los datos, vamos a realizar un análisis de componentes principales(el análisis completo está en el apartado 1.3 del anexo, donde está todo explicado,lo hemos redactado ahí por cuestiones de reducir el tamaño del trabajo).

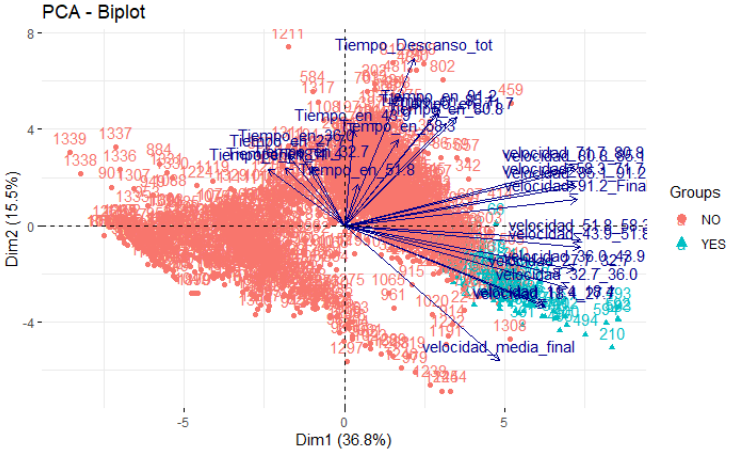
Hemos realizado el PCA, con las velocidades medias por etapa, y los tiempos de descanso en cada punto de control, para determinar si su valor en ciertos puntos a lo largo de la carrera, son más determinantes en la distinción de la élite de los corredores del resto.

BEAR



En este gráfico de scores coloreado según si el corredor pertenece a los 10 mejores participantes de su año(azul) se pueden ver los pesos de las variables del tiempo de espera y las velocidades medias.
Es difícil ver los nombres de las variables debido a la aglomeración, pero en el gráfico de pesos y correlaciones con las componentes principales en el anexo puede verse más claramente.
Vemos que las velocidades medias de todos los puntos de control de la carrera contribuyen a separar entre estos dos grupos, por lo que una mayor velocidad media en las etapas a lo largo de la carrera lógicamente determinaría si un corredor es del top 10 o no, especialmente a partir de las etapas medias ya que los corredores menos preparados sufren una disminución de velocidad media a partir de ese punto.

CLOCKWISE HARD ROCK



Observando este gráfico de los pesos, vemos como en esta carrera, las velocidades medias y los tiempos de descanso determinan la separación de los corredores de élite y el resto. Primero,vemos como la variabilidad de los corredores de elite hacia la esquina inferior derecha, viene determinada por las velocidades medias entre las millas 11.4 y 43.9 por lo que estas son las más cruciales en la carrera para determinar un mejor tiempo final y cuanto mayores las velocidades en este tramo, más hacia el grupo de corredores de élite se distribuyen los scores. También la velocidad media final está negativamente correlacionada con los tiempos de descanso en algunas de las etapas iniciales ,es por ello que algunos de los corredores de élite presentan mayores velocidades medias para dichas etapas así como menor tiempo de descanso, no obstante el resto de corredores tienen una velocidad media más reducida así como más tiempo de descanso ,indicando así una diferencia entre los corredores de distintos niveles.

Al mismo tiempo la variabilidad de los scores de corredores hacia la esquina superior izquierda es determinada por los tiempos de descanso en los puntos de control entre las millas 18.4-36 por lo que podemos deducir que cuanto más descanso realice en las etapas iniciales, este tendrá una menor velocidad media.

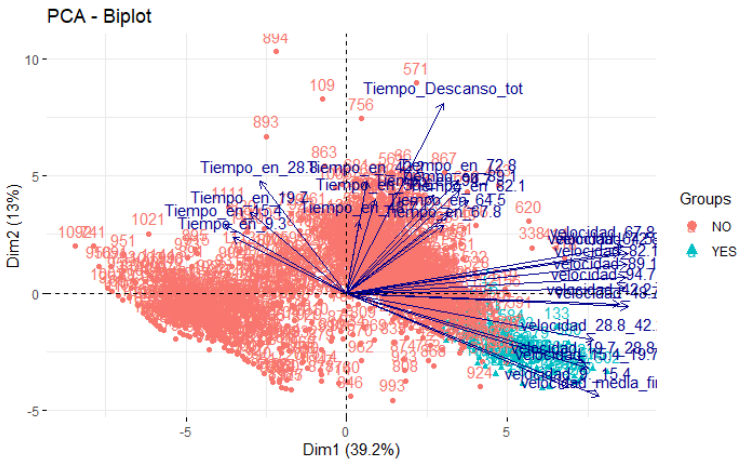
COUNTERCLOCKWISE HARD ROCK

Ya que la diferencia entre el circuito clockwise y counterclockwise es la dirección, y algunos puntos de control, ambas carreras presentan un patrón similar

Vemos cómo al igual que en el circuito clockwise, la distribución de los scores de los corredores elite viene explicada por las velocidades medias en la primera mitad de la carrera, entre el inicio y la milla 42.2, esto puede ser debido a la dureza inicial de la carrera, por lo que una mayor velocidad media en estas etapas acerca más a los corredores al grupo de élite.

Al igual que en clockwise las proyecciones de las velocidades que contribuyen a la explicación de la distribución de los corredores de élite, están correlacionados negativamente con algunos tiempos de descanso al estar en ángulo de 180 grados.

En comparación con la carrera clockwise, los tiempos de descanso tienen más peso en la variabilidad, vemos que estos descansos son los que ocurren entre las millas 9.3 y 28.8 por lo que un mayor tiempo de descanso en estas millas aleja los scores del grupo de corredores de élite



¿Es posible predecir si un corredor va a abandonar o proseguir su carrera dependiendo de su actuación en ciertas etapas?

Queremos determinar si con los datos de un corredor de los tiempos entrada en puntos de control, velocidades medias, y tiempos de descanso, podemos predecir si abandona o no la carrera, hemos elegido estos datos para las etapas entre las 30 y 60 millas ya durante la primera parte los corredores aún están frescos y no es posible determinar aún si los corredores van a abandonar o no. Partimos los datos para cada carrera en 80% datos de entrenamiento y 20% datos test, y con los datos de entrenamiento usando el estado final del corredor(abandona/termina) como variable respuesta y los datos numéricos que describen como lo hace el corredor en esas etapas creamos un modelo que comprobamos con los datos test, sin embargo no obtenemos buenos resultados.

En el apartado 1.4 del anexo mostramos los resultados que no son concluyentes, por lo que con los datos que tenemos no podemos predecir el estado del corredor con estos datos.

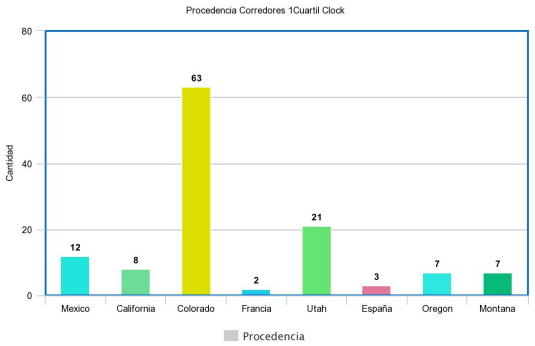
Estudiar la Internacionalidad de la carrera.

Estas carreras son pruebas realmente duras donde cada atleta demuestra lo mejor de sí mismo, todas ellas son de gran duración y con una orografía de terreno muy dura, por estos motivos estos circuitos sólo son completados por corredores de muy alto nivel.

Es por ello que vamos a investigar las 8 procedencias más comunes dentro del primer cuartil de corredores de cada carrera con el fin de poder relacionarlo con la conclusión del proyecto.

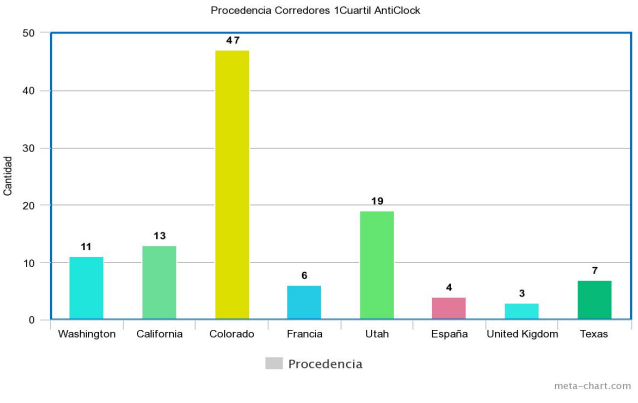
Hardrock clockwise

En el caso de HardRock Clockwise podemos ver como se trata de una carrera con cierta intencionalidad. Entre las 8 procedencias más comunes de los corredores del primer cuartil destacan Colorado y UTAH, no obstante también tenemos corredores de México, Francia, y España entre otros.



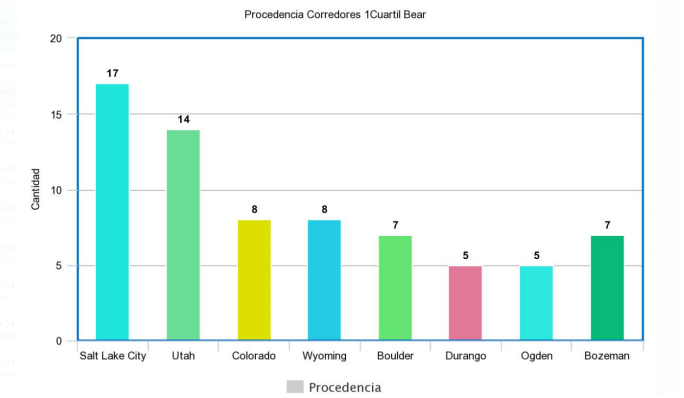
Hardrock counter-Clockwise

Tras observar el gráfico de barras se puede determinar que el circuit de Anticlockwise es también de ámbito internacional, sin embargo como ocurría en el circuito de Clockwise las localizaciones más frecuentes son Colorado , California y Utah, teniendo una pequeña representación países Europeos como son España, Inglaterra y Francia.



The bear

Sorprendentemente, a diferencia de los circuitos de circuitos de Hard Rock , bear es el circuito con menor internacionalidad, las localizaciones más frecuentes entre los corredores son Salt Lake City , Utah, Wyoming, Colorado… Estas procedencias son muy frecuentes debido a la cercanía al circuito de Bear, siendo Salt Lake City y Ogden las más cercanas (ambas situándose en UTAH).



Velocidades medias y los tiempos de espera de las etapas en función de la orografía de la etapa.

Aquí tenemos una visualización del rendimiento promedio de los mejores corredores de cada carrera.

Esto lo hemos conseguido parseando un archivo GPX para cada carrera, que contiene un esquema XML con un registro GPS de puntos cercanos y equidistantes entre sí con sus coordenadas geográficas, su elevación sobre el nivel del mar, y la fecha y hora en la que se ha registrado. Esto último es útil para calcular velocidades exactas en cada punto, pero no lo hemos utilizado puesto que el archivo lo hemos obtenido de páginas como Alltrails o Wikiloc, y por tanto no podíamos ver las velocidades de nuestros corredores. Esa ha sido nuestra tarea principal en este apartado.

Para empezar hemos utilizado datos obtenidos mediante web scraping con R de OpenSplitTime. Para cada carrera necesitábamos dos ficheros de datos diferentes, uno con los 'Splits' de cada carrera (es decir, información de cada punto de control, coordenadas, altitud y distancia recorrida, a las que les hemos agregado datos calculados) y otro con información de cada corredor y las horas por las que ha pasado por cada punto de control.

Mediante pandas de Python y funciones que hemos diseñado, hemos añadido a la base de puntos de control la pendiente (%), y la distancia y la elevación desde el punto de control anterior, para posteriormente calcular las velocidades promedio de cada corredor para cada punto de control.

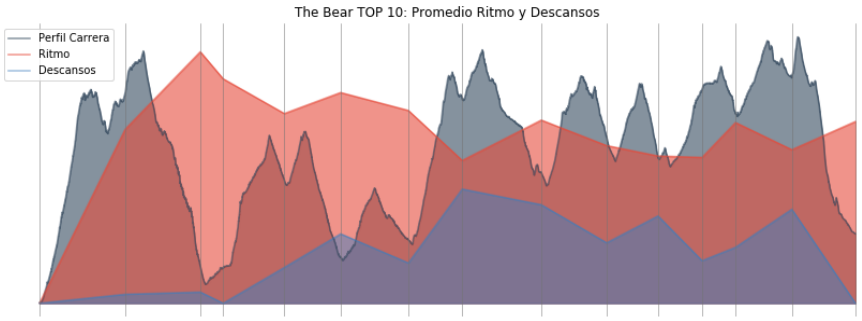
Hemos representado estas velocidades promedio y tiempos de descanso para cada punto de control de los campeones de cada carrera encima del perfil de elevación del circuito, para poder visualizar e interpretar correctamente el rendimiento de estos.

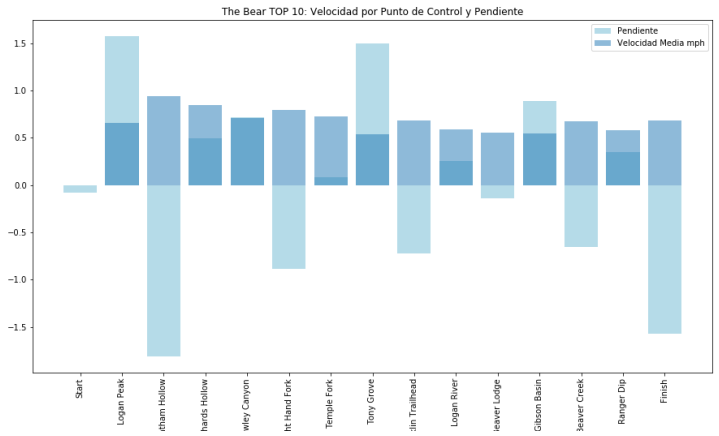
Nota: para poder representar rangos de datos tan diferentes en una misma gráfica hemos tenido que normalizar los valores a rangos de datos más compatibles, y en el caso de las pendientes hemos normalizado y centralizado los valores mediante un escalado estándar utilizando la desviación típica y la media, para poder visualizar los valores negativos correctamente.

En todos los circuitos podemos ver un comportamiento similar y predecible de los corredores, ya que vemos los mayores incrementos de velocidad cuando la pendiente es negativa, y disminuciones de esta en las cuestas. No obstante vemos que a pesar de haber velocidades más altas y más bajas los corredores intentan mantener una estrategia con un ritmo bastante constante (en la medida de lo posible, adaptando la velocidad a los cambios del terreno), ya que no hay grandes sprints ni descensos de velocidad excesivos. Esto se debe a que el cuerpo resiste mejor un ritmo constante para esfuerzos prolongados que un ritmo con grandes variaciones. Además en bajadas prolongadas aumenta considerablemente el riesgo de lesión de rodillas, articulaciones y espalda, por lo que los corredores deben de ir a una velocidad dentro de sus límites para no tener que retirarse antes de acabar. Esto lo podemos ver de forma clara en los gráficos de barras.

Esta estrategia de carrera adaptable al terreno está considerada la más real, aunque es de las menos estudiadas científicamente.

THE BEAR



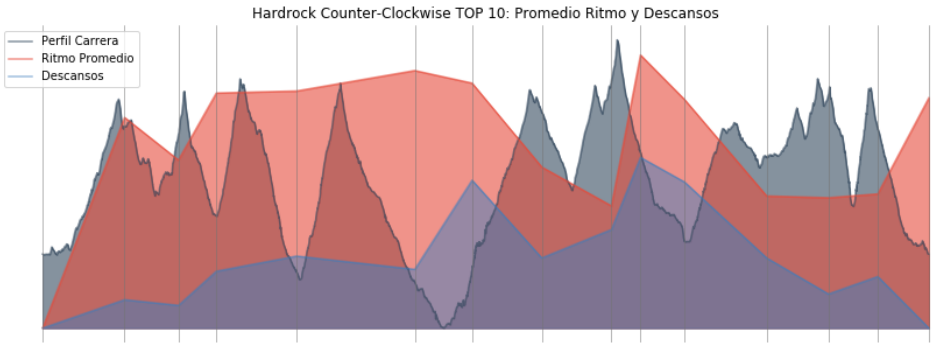


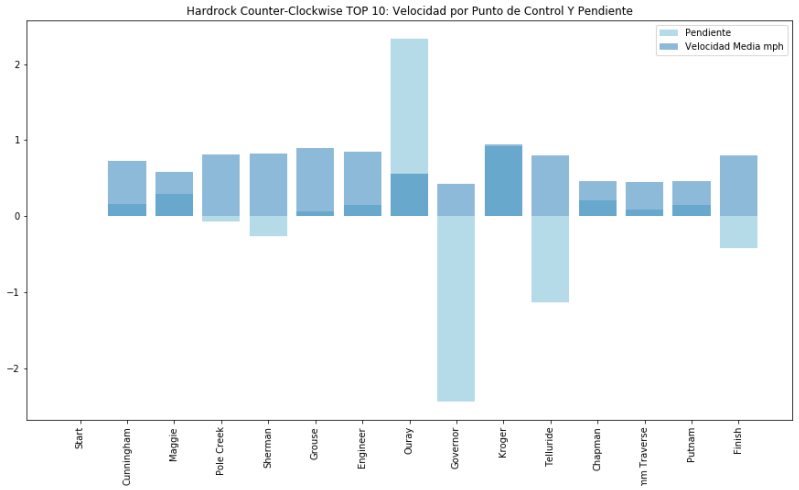
Para esta carrera vemos el mayor pico de velocidad entre el primer y segundo punto de control (Logan Peak y Leatham Hollow), ya que es el tramo con mayor desnivel negativo.
Por otro lado, la velocidad más lenta es en el tramo entre el sexto y séptimo punto de control (Temple Fork y Tony Grove) y vemos que este tramo coincide con la segunda mayor cuesta de la carrera. La cuesta más pronunciada se produce entre la salida y el primer punto de control, pero podemos comprobar que en este caso la velocidad no es tan baja, ya que el cansancio aún no ha empezado a hacer efecto en los corredores.

Por lo demás, y a parte de ser un terreno con muchos altibajos, podemos observar que el ritmo promedio para los mejores corredores no sufre grandes variaciones durante el total de la carrera, ya que siguen una estrategia de ritmo adaptable al terreno.

Por el contrario, vemos que los tiempos de descanso van aumentando a medida que avanza la carrera, como es de esperar. Observamos que el punto de control en el que los corredores de élite pasan más tiempo descansando es en el número 7 (Tony Grove), que es el punto de control donde acaba la cuesta mencionada antes.

HARDROCK COUNTER-CLOCKWISE



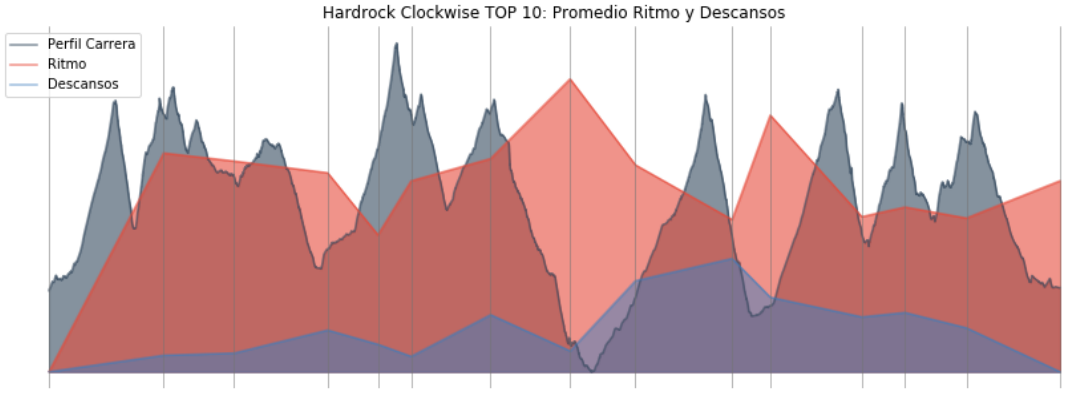


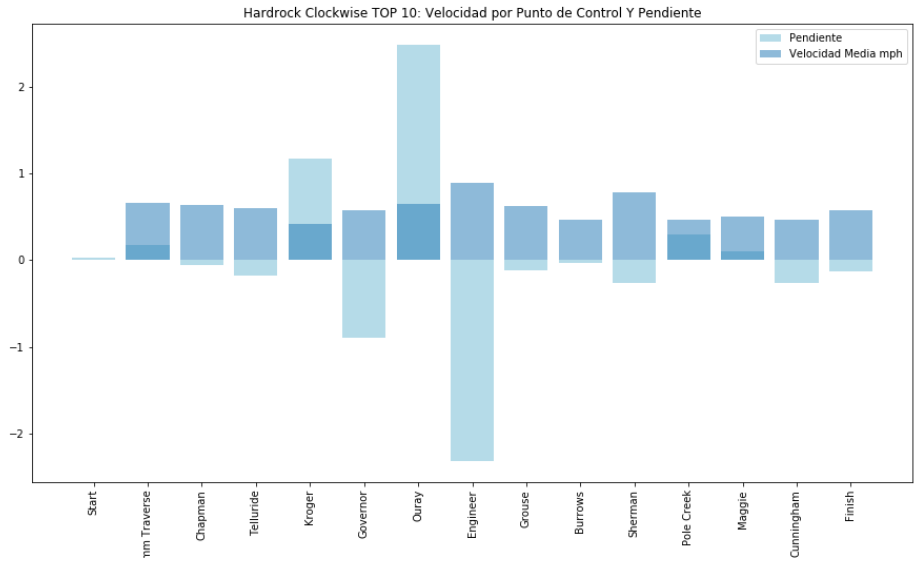
Aquí vemos algo muy parecido. Lo que más llama la atención a simple vista es el tramo entre los puntos de control 3 y 6 (Pole Creek y Engineer), ya que el ritmo promedio no varía prácticamente. Obviamente en la realidad esto no es así, tiene una explicación. Si nos fijamos en los 3 tramos en los que se divide, vemos que cada uno de ellos tiene una zona de bajada y otra de subida, o viceversa, similares en cuanto a distancia y pendiente, por lo que la velocidad promedio no se ve afectada.

Otra cosa que nos llama la atención es una variación de velocidad considerable entre los puntos de control 6 y 10, ya que en esta zona se encuentra el pico más alto de la carrera y es imposible mantener el ritmo constante. Vemos que hasta que se llega al pico la velocidad

promedio disminuye hasta convertirse en la zona con el ritmo más lento de la carrera, pero una vez se pasa este pico la velocidad aumenta en gran medida, siendo la zona más rápida de la carrera. Curiosamente es después de esta bajada el tramo donde los corredores suelen detenerse más tiempo a descansar, y no en la cima, como cabría esperar. Esto se puede deber a la gran exigencia de ese pico, tanto en la subida como en la bajada, ya que la subida es un gran desafío a la resistencia de los corredores, y en la bajada se ponen a prueba las articulaciones y la espalda, ya que es un tramo donde sufren mucho. Podemos deducir por tanto, por qué los corredores necesitan un breve respiro al acabar ese tramo.

HARDROCK CLOCKWISE





Este es el mismo circuito que el anterior, pero en sentido contrario, por lo que podemos esperar que rendimiento de los ultramaratonianos sea similar. En general vemos un ritmo bastante constante, ya que la mayoría de puntos de control tiene tramo de subida y tramo de bajada. Lo más destacable de este circuito es la subida que se produce entre los tramos 3 y 4, que reduce el ritmo a un mínimo; y la gran bajada que se produce entre los tramos 6 y 7, en la que se produce el pico más alto de velocidad media. Después llega otra cima parecida a la del circuito anterior, con una subida en la que los corredores disminuyen el ritmo, y un descenso en el que estos aprovechan para acelerar. Vemos que cuando acaba esta bajada es el tramo

donde los corredores pasan más tiempo, por lo que vemos un comportamiento similar al de Hardrock Counterclockwise, y podríamos pensar que esta es una cima de gran exigencia física.

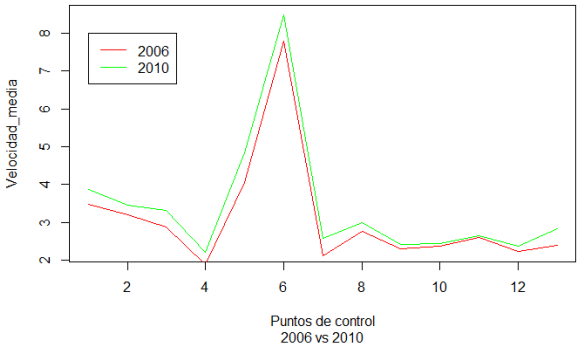
Efecto de *la climatología en las velocidades y tiempos de espera*

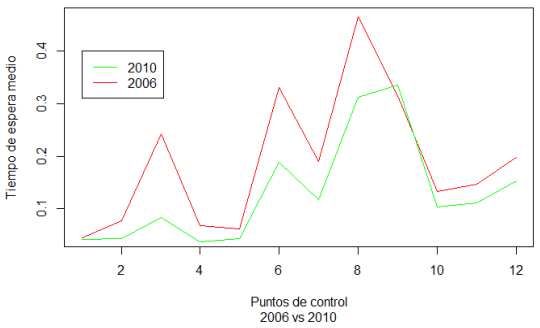
Respecto a esta pregunta se ha de dividir en 2 partes, ya que el tiempo va a afectar de 2 formas diferentes, 1º si hay condiciones desfavorables antes de la carrera (suelo mojado) o si hay condiciones favorables/desfavorables durante la carrera.

Las precipitaciones los días previos a las carreras, se ve en el estudio que en ninguna de las ediciones de las distintas carreras han sucedido las suficientes condiciones climáticas adversas (lluvias) para que suponga un problema para los participantes a la hora de realizar una carrera. Todo esto era de esperar ya que estas carreras tienen décadas de antigüedad y se hacen estratégicamente en épocas y meses en los que las condiciones meteorológicas influyen lo mínimo posible.

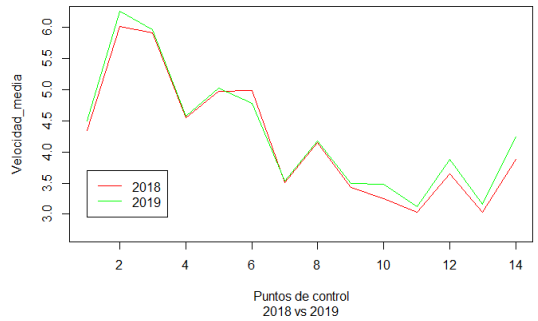
La climatología en la carrera, tras coger un total de 11 muestras entre las 3 carreras (1 muestra es comparar una edición con otra, claramente con ciertas particularidades diferentes cada una de ellas, **APARTADO 1.1 ANEXO**), se ha podido constatar que a cuanto más temperatura relativa se exponga a un corredor, de media su velocidad media se verá reducida y sus tiempos de espera entre puntos de control (descansos) serán mayores que en condiciones climáticas favorables (con una temperatura más reducida y sin precipitaciones).

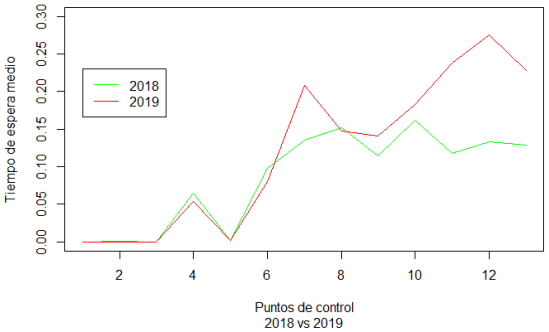
Un claro ejemplo de este hecho es en la carrera de Hardrock-clockwise donde la edición de 2006 hizo más calor que en la edición de 2010, la velocidad media disminuyó de forma considerable y los tiempos de espera en los diferentes puntos de control aumentaron considerablemente también.





En cambio las precipitaciones (a pesar de no tratar con demasiados datos de este tipo), se ha podido observar un patrón respecto a las ediciones estudiadas, en todas las carreras donde hay una cierta precipitación a priori sus corredores no se ven tan afectados en su velocidad media entre etapas, pero en cambio sus tiempos de espera aumentan significativamente entre etapa y etapa (descansos) en la mayoría de ocasiones. Un claro ejemplo es la carrera de The Bear, la comparación de las ediciones de 2019 (la cual tuvo precipitaciones durante toda la carrera), con la edición de 2018 que no tuvo condiciones climatológicas adversas.





Eso nos lleva a concluir que la temperatura relativa es la mayor responsable en la reducción de las velocidades medias entre ediciones y del aumento de los tiempos de espera entre puntos de control, aunque en esta última también puede tener un gran peso la cantidad de precipitaciones que sucedan en la edición, si hay un gran número de precipitaciones de media los corredores pasarán más tiempo en los puntos de control.

¿Como la edad o el género influyen en los resultados?

En esta sección mostraremos el análisis final de las variables que hemos seleccionado y que, como norma general, marcan la diferencia en los tiempos de las carreras, a través de las distintas carreras y veremos cómo y en qué grado afectan.

HARDROCK COUNTER-CLOCKWISE

En primer lugar para la carrera de Counterclockwise obtenemos como aproximadamente de los 180 corredores , 160 de ellos son hombres y 20 mujeres. (Fig 1.2.1)

Debido a la gran diferencia de atletas según sexo podemos pensar que en el caso de las mujeres los resultados no sean tan significativos ya que tratamos con una muestra mucho menor. No obstante vamos a comparar los Tiempos finales según género.

Como podemos apreciar el tiempo medio de las mujeres es ligeramente inferior al de los hombres (indicando una menor marca) , sin embargo en el caso de los atletas más extremo, la diferencia entre ambos sexos es mínima y no obstante en el caso del mejor atleta vemos como el atleta más rápido es un hombre con una diferencia de 4 horas respecto a la primera mujer. (Fig 1.2.2)

Otro factor que hemos pensado que sería muy relevante estudiar ya que es una de las preguntas más realizada entre los runners es el efecto de la edad.

Si observamos la (Fig 1.2.3) podemos visualizar una relación muy plausible, en términos de la media, a menor edad se obtiene una mejor marca, no obstante hay casos de corredores con rangos de edades de 25-45 años con tiempos inferiores a los corredores más jóvenes, también resulta interesante que para los corredores muy jóvenes o seniors el rango de tiempos es mucho menor que en el caso de corredores de 25-45 años.

En el caso de las velocidades medias tenemos un comportamiento similar que en el de los tiempos finales, las velocidades medias más altas ocurren para los atletas de menor edad, esta

media baja con el aumento de edad, no obstante existen corredores de 25-45 años con velocidades muy superiores al resto de los participantes.(Figura 1.2.4)

HARDROCK CLOCKWISE

A continuación realizamos el mismo procedimiento para el caso de clockwise.

Como era de esperar y basado en la cantidad de hombres/mujeres del circuito de counterclockwise podemos ver una relación similar, tenemos aproximadamente 170 hombres y 20 mujeres en el primer cuartil de la misma. Figura (1.3.5)

Comprobamos como para clockwise los tiempos finales respecto al género son bastante diferentes respecto a anticlockwise, en este circuito los hombres y las mujeres tienen aproximadamente los mismos tiempos como media , y los 'peores' corredores de este primer cuartil tienen marcas casi idénticas, sin embargo en el caso de los mejores corredores si que podemos afirmar que el corredor más rápido es un hombre, y este es aproximadamente 6 horas más rápido que la primera mujer . Figura (1.3.6)

En el caso del tiempo final según edad se aprecia un comportamiento similar que al circuito anteriormente descrito, como norma general cuanto más jóvenes son los atletas mejores marcas obtienen y a medida que los rangos de edades aumentan vemos menos diferencias entre los distintos tiempos de los corredores. Figura (1.3.7)

Las velocidades medias según edad también tienen una relación con el circuito de counterclockwise, a menor edad mayor velocidad media, y mayor el rango de velocidades entre los corredores. Esto es apreciable en la figura (1.3.8)

THE BEAR

Realizamos el mismo mini análisis al circuito de Bear, el primer cuartil de esta carrera está formada por aproximadamente la misma cantidad de hombres/mujeres que el resto de circuitos, en este cuartil tenemos aproximadamente 180 hombres y 20 mujeres.(Figura= 1.3.1)

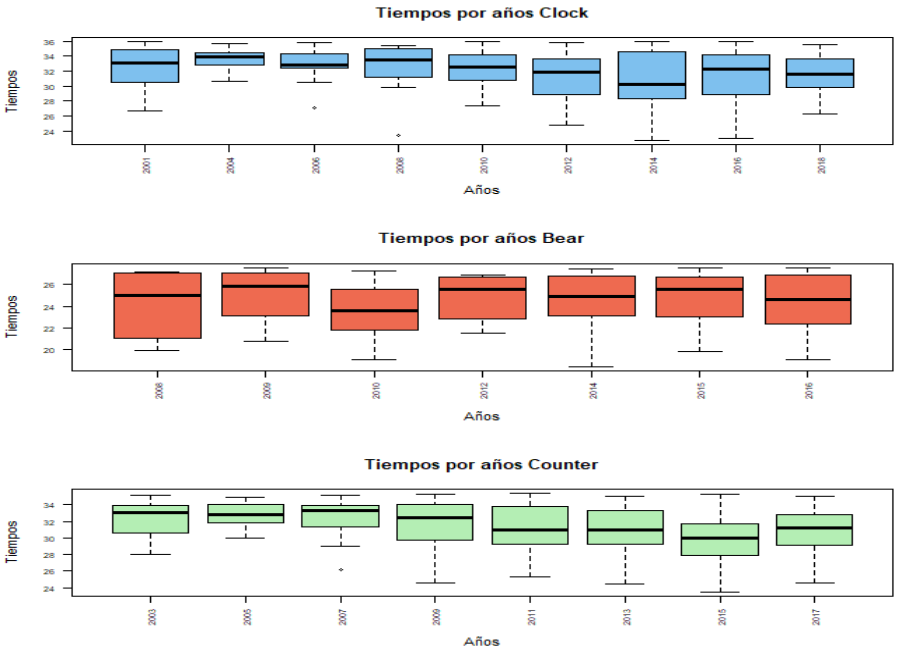
Respecto a los tiempos finales según sexo, volvemos a ver un patrón parecido al caso del circuito clockwise, la media de los hombres es ligeramente inferior que la de las mujeres y el hombre más rápido es aproximadamente 3 horas mas rápido que la mujer más rápida, también cabe mencionar que el rango de tiempos en el caso de las mujeres es mucho menor, todo esto es causado debido a que están menos representadas en este cuartil de tiempos.(Figura = 1.3.2)

Si analizamos los datos entorno a tiempo final y edades podemos ver cómo sigue el mismo patrón, en general a menor edad mejor el tiempo medio del corredor, al mismo tiempo es

importante mencionar que existen corredores del rango 25-50 más veloces que aquellos más jóvenes. Resulta curioso mencionar que en el circuito de Bear el atleta más rápido se encuentra en el rango de 46-50 años.(Figura= 1.3.3)

Para finalizar con este mini análisis explicaremos la relación entre las velocidades medias y las edades, como hemos podido apreciar en el resto de circuitos los corredores de menor edad tienen la mayor velocidad media, según las edades aumentan las velocidades medias disminuyen así como los rangos.(Figura=1.3.4)

Analizar el cambio en los tiempos en cuestión de años



Al observar los boxplots apreciamos como por norma general a medida que pasan los años el rendimiento medio y el rendimiento de los atletas más extremos es mejor , estas cuestiones serán explicadas en la conclusión del proyecto.

Lecciones aprendidas para mis futuros proyectos en Ciencia de Datos.

En este proyecto todos los integrantes del grupo han tenido que enfrentarse a muchas dificultades y saber cómo solucionarlas en tiempos reducidos, gracias a eso se ha aprendido mucho, sobre todo cómo buscar información, tanto técnica como sobre el tema en cuestión, en internet, y aplicar eso al estudio particular, además de mejorar en organización y eficiencia de trabajo.

También se ha aprendido a filtrar y seleccionar información precisa y necesaria y eliminar todo lo que no nos interesa, aparte de los conocimientos técnicos adquiridos tanto en R como en Python y web scraping, esto será muy beneficioso para próximos proyectos.

5. Conclusiones

En conclusión podemos apreciar cómo en estas carreras se presentan una mayor cantidad de hombres que de mujeres, la mayoría de estos corredores se encuentran en el rango de los 25-45 años, entendemos que entre este rango de edades se encuentra la edad óptima para realizar actividades de largas distancias debido a su experiencia. Estos participantes siguen mejores estrategias en términos de uso de energía, como por ejemplo llevar un ritmo mínimamente constante y adaptado a las variaciones del terreno, sin acelerar demasiado en las bajadas y sin frenar excesivamente en las subidas, a fin de controlar y mantener la homeóstasis del cuerpo, evitar tempranas lesiones durante el trayecto y llevar una buena economía de la carrera.

Al mismo tiempo debido a la gran experiencia de los diversos corredores de dichas edades y los diferentes tipos de entrenamientos usados por cada uno de ellos podemos entender la razón por la cual los rangos de tiempos finales son tan diferentes para este colectivo, al mismo tiempo un gran ejemplo para poder demostrar esto es que la mayoría de corredores del primer cuartil son de localidades muy cercanas a donde se realizan las competiciones. Con esto podemos intuir que estos atletas entrenan en condiciones climatológicas muy similares y posiblemente hayan podido practicar estos circuitos con anterioridad, y de este modo han podido realizar un entrenamiento más específico para mejorar su rendimiento.

Es importante mencionar como hemos observado que a mayor temperatura relativa los corredores necesitan más tiempo de descanso y tienen menor velocidad de media. Esto es un claro ejemplo de una medida usada por los susodichos para combatir la deshidratación, y en el caso de precipitaciones, los tiempos medios de descanso son más elevados, ya que al correr con malas condiciones climatológicas los deportistas tendrán que estar más atentos al entorno, por lo que acabarán con una mayor fatiga.

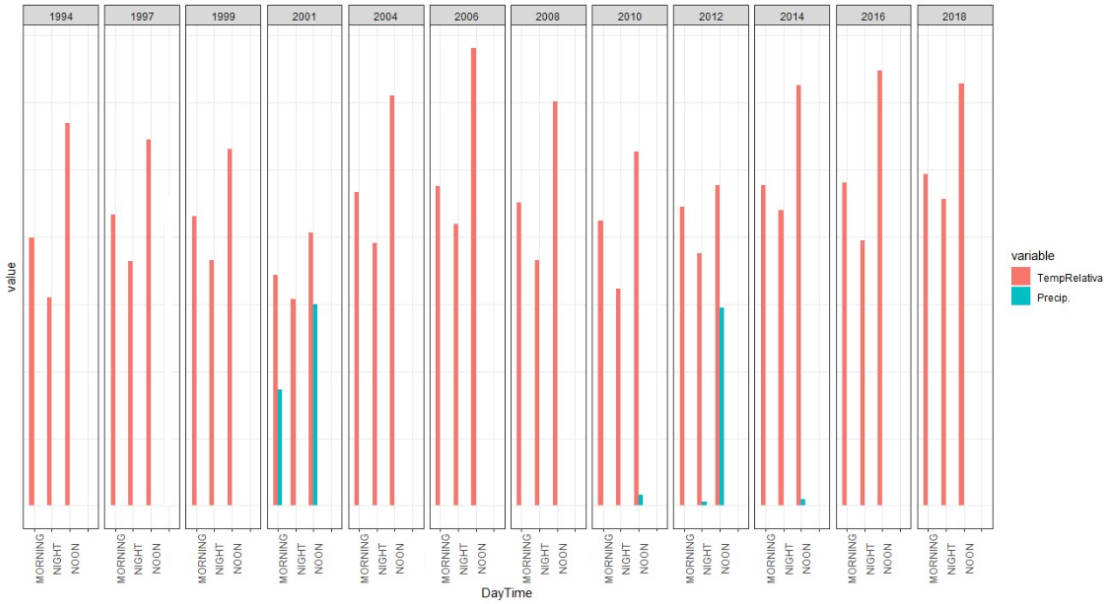
Para finalizar podemos afirmar como los tiempos medios a lo largo de los años han ido disminuyendo, esto es un claro ejemplo de la evolución científica en términos de entrenamientos más eficientes, así como la importancia de la suplementación y el correcto calzado y equipamiento con el fin de reducir las probabilidades de lesión y mejorar la recuperación.

ANEXOS

APARTADO 1.1

RELACIONES TEMPERATURA Y VELOCIDAD MEDIA

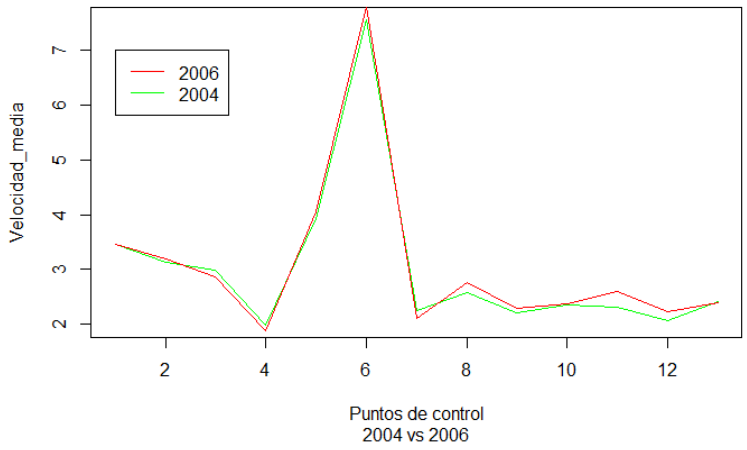
HARDROCK CLOCKWISE:



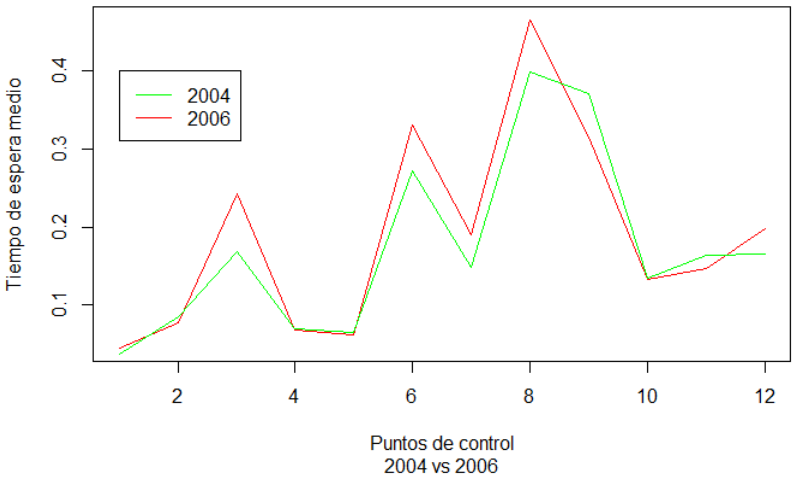
Al ver los gráficos de precipitaciones y de temperaturas relativas, se quiere estudiar cómo afecta las condiciones meteorológicas a los 50 mejores corredores de Hard Rock Clockwise de cada edición.

2006-2004

Primeramente se comparan los años de 2004 y de 2006 (donde las temperaturas de 2006 aumentan ligeramente respecto a 2004), como se puede observar a priori un ligero aumento de la temperatura no afecta a la velocidad media de estos corredores.

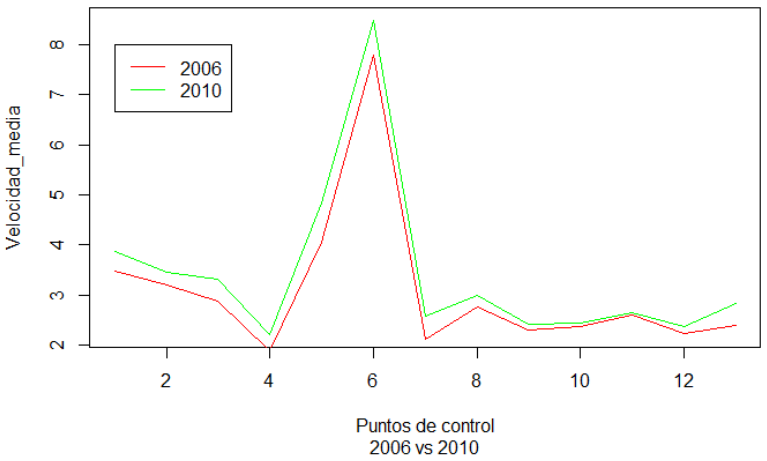


Pero en cambio si aumenta un poco la temperatura (tal y como pasa en 2006), los corredores de media descansan bastante más tiempo en los puntos de control.

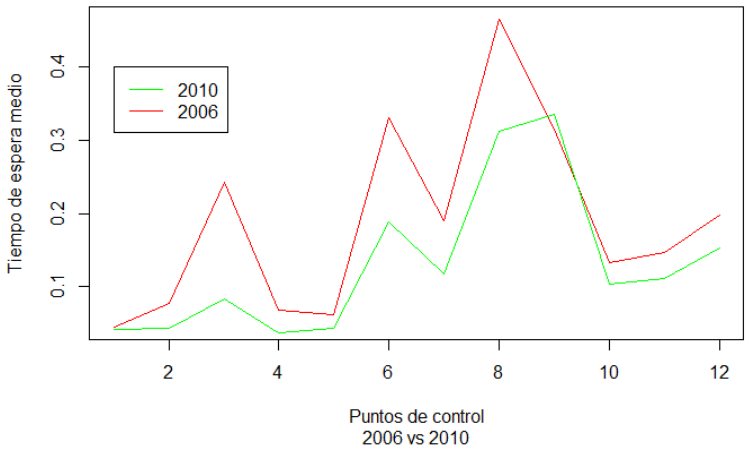


2006-2010

Ahora comparando los datos de velocidad media de 2006 con la de 2010 (dónde 2010 hizo menos calor), hay una clara disminución de la velocidad media de los corredores.

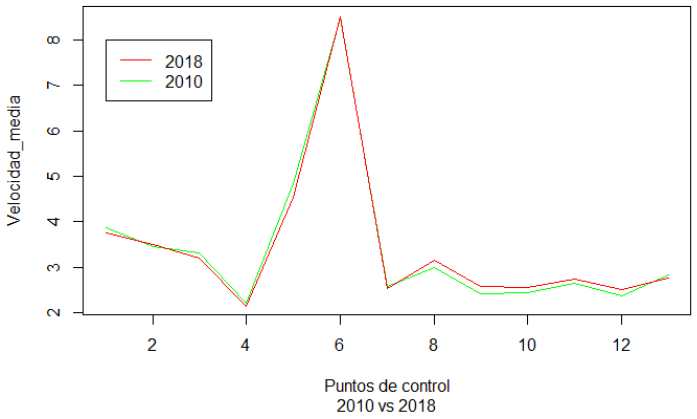


Esta dinámica continua claramente en los tiempos de espera, donde en la edición donde hizo más calor aumentan significativamente los tiempos de espera respecto a la edición de 2010.

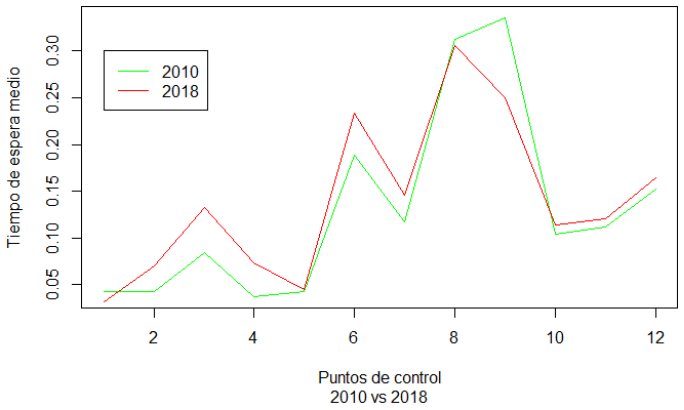


2010-2018

Comparando ahora las velocidades medias de 2010 y 2018 la no afecta en gran medida a los corredores, aunque en 2010 los tiempos son ligeramente más rápidos.

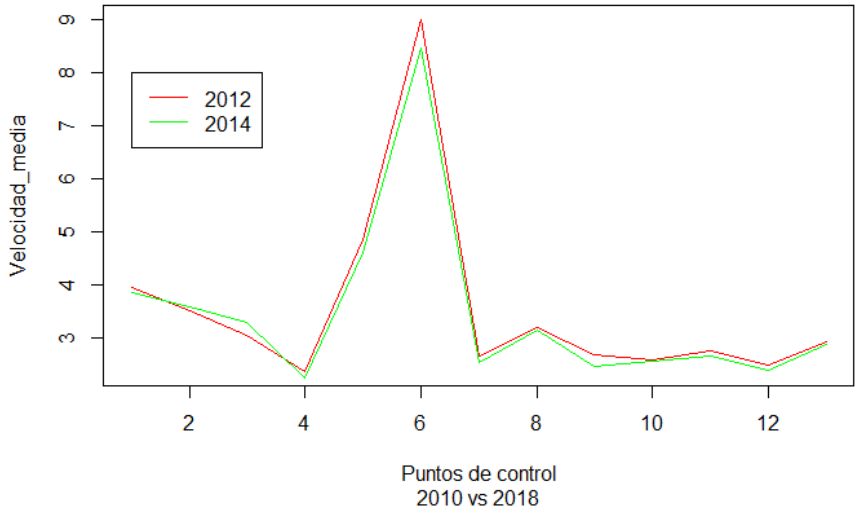


En cuanto a los tiempos de espera se puede observar una clara diferencia entre 2010 (donde hizo menos calor) y 2018, pero esta tendencia se rompe en el noveno punto de control, debido a que en esas horas hubo una ligera precipitación (en la edición de 2010) que al parecer provocó que los corredores tuvieran que descansar más tiempo.

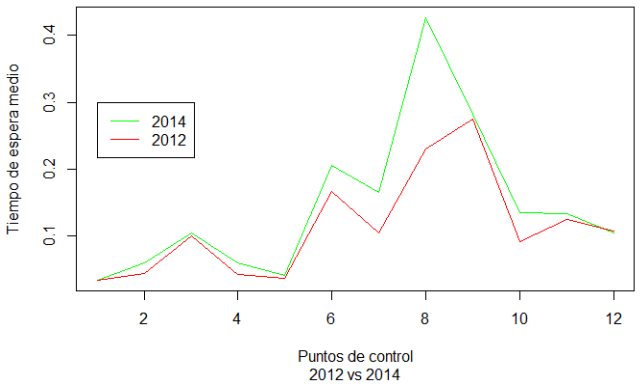


2012-2014

Viendo los años 2012 y 2014, donde en 2012 hubieron precipitaciones y en 2014 temperaturas elevadas, se puede observar que las temperaturas elevadas afectaron en cierta manera a la velocidad media de los corredores, haciendo que estas aumentarán.

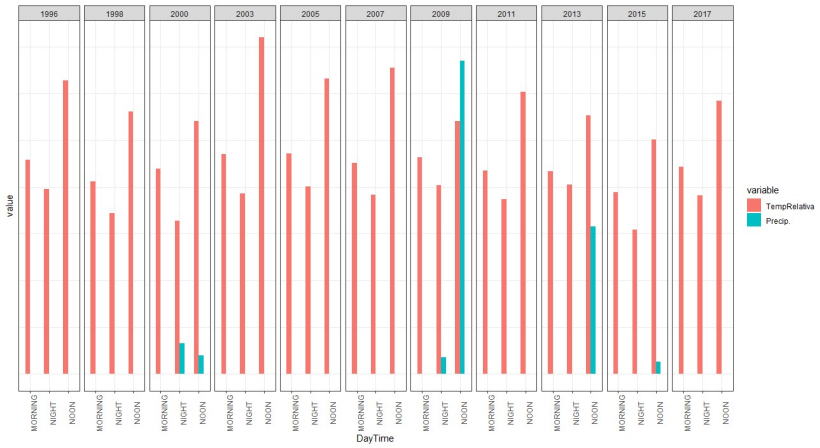


En cambio viendo los tiempos de espera entre puntos de control se observa que la edición de 2014 la más calurosa de las 2 que se están estudiando, sus participantes de media pararon mucho más tiempo en los puntos de control.



Pero en los tiempos de espera se puede ver que la temperatura afectó en mayor medida a los corredores durante las otras etapas que la lluvia de la edición de 2012, es decir, debido a la temperatura los corredores de la edición de 2014 estuvieron más tiempo descansando.

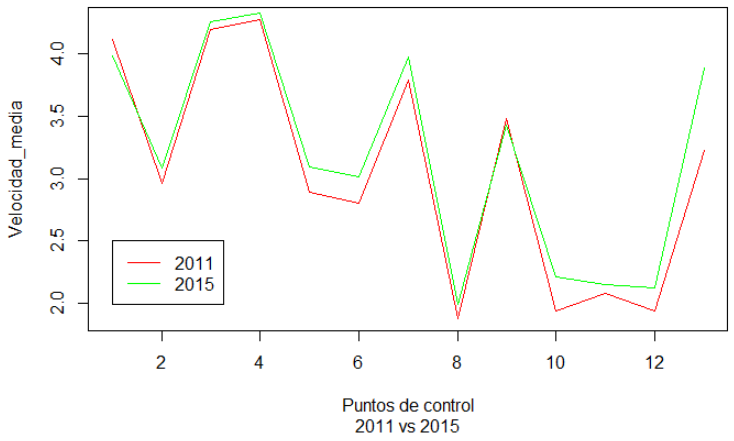
HARDROCK COUNTER-CLOCKWISE



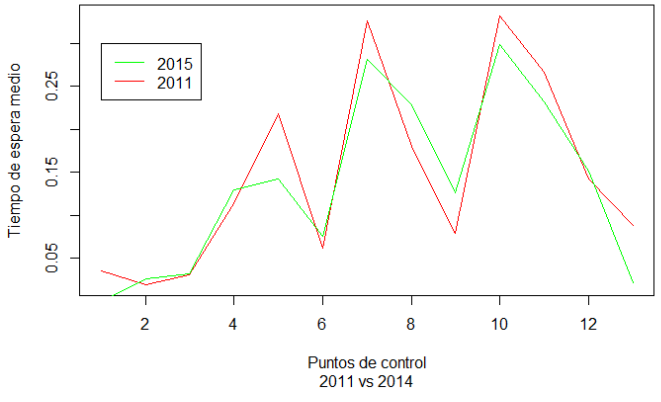
Al ver los gráficos de precipitaciones y de temperaturas relativas, se quiere estudiar cómo afecta las condiciones meteorológicas a los 50 mejores corredores de Hardrock Counter-Clockwise de cada edición.

2011-2015

Primeramente se comparan los años de 2011 y de 2015 (donde las temperaturas de 2011 aumentan respecto a 2015), como se puede observar a priori el aumento de temperatura ha provocado que los corredores reduzcan su velocidad media respecto a la edición de 2015.



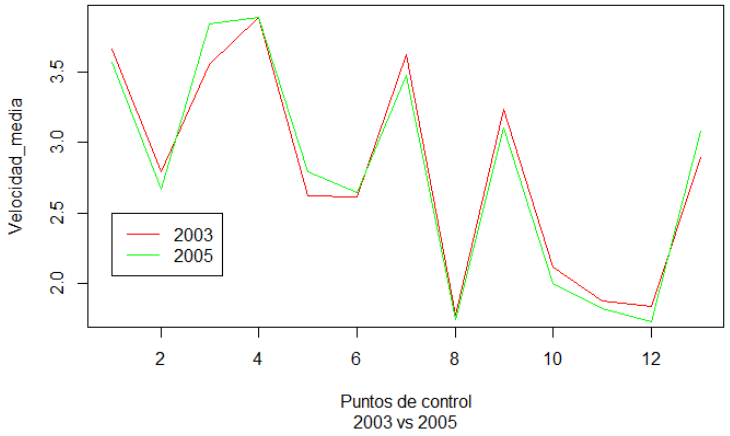
Esta tendencia continúa en los de espera medios, donde en la edición de 2011 debido a las altas temperaturas, hay un mayor tiempo de espera (descansos) respecto a a la edición de 2015 donde hubo menos calor.



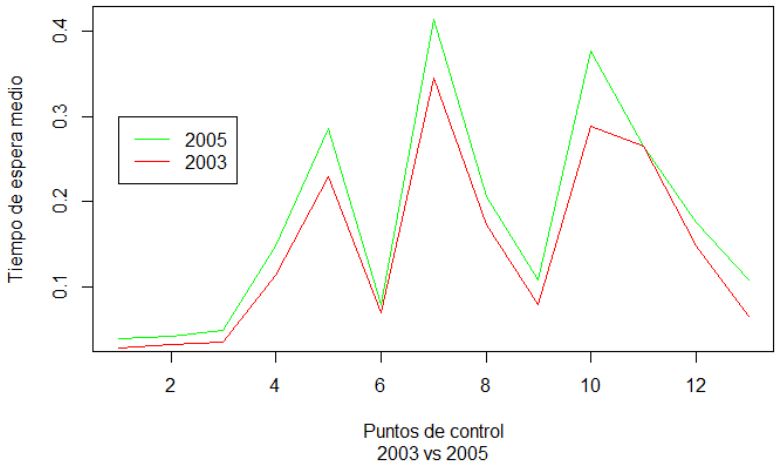
2003-2005

En esta gráfica se van a representar 2 ediciones, las cuales hubieron por su parte temperaturas bastante similares, por tal de ver si al aumentar un poco la temperatura relativa aumenta por su parte la velocidad media de los corredores de dicha edición.

Como se puede ver en la gráfico en los primero compases de la carrera (las que transcurren al medio día) los corredores de la edición de 2003 tienen una velocidad media por etapa menor a la edición de 2005, este hecho más tarde se va proporcionando en función de conforme van pasando los puntos de control.

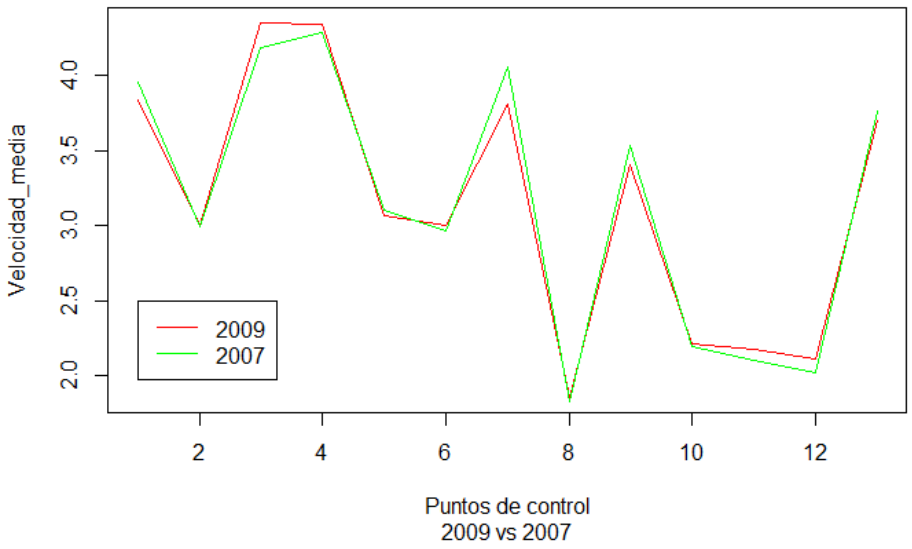


Curiosamente, en los tiempos de espera entre etapas, las etapas de espera de 2003 son mucho más pequeños de media que los tiempos de espera de la edición de 2005

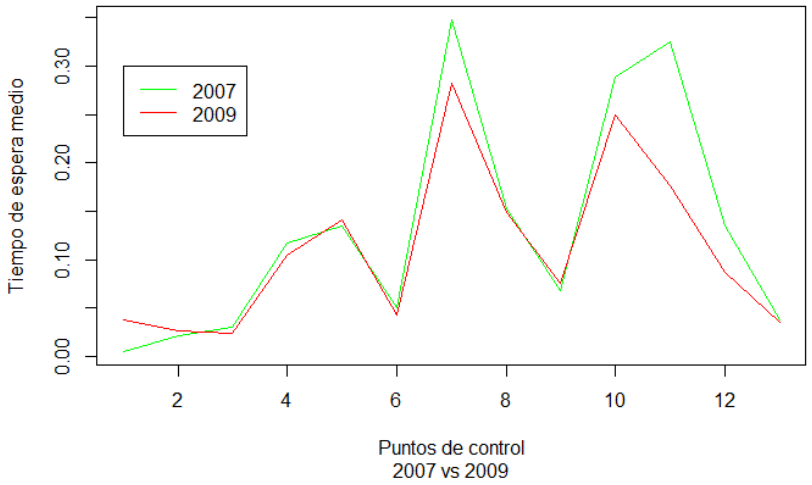


2007-2009

Comparando los años 2009 y 2007, se puede ver que la temperatura del edición 2007 son superiores a la temperaturas de la edición de 2009, pero en esta última hay precipitaciones en la franja del mediodía. Como se puede observar los corredores de la edición de 2009 cuando Coincide en esa franja sus velocidades medias son más reducidas respecto a la edición de 2007.

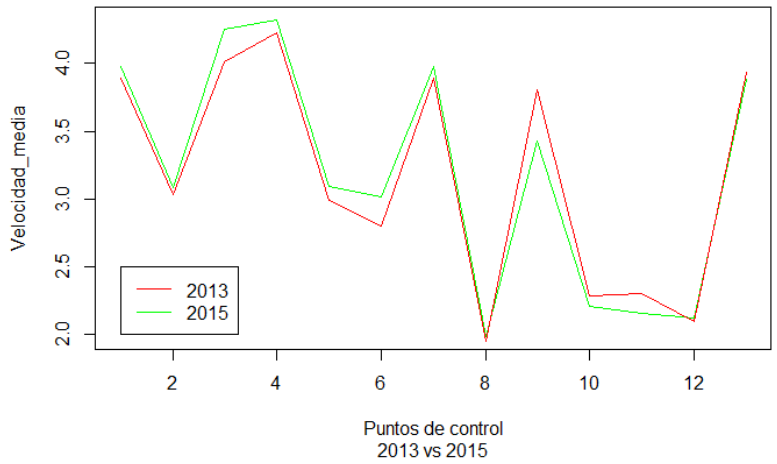


En cambio en los tiempos de espera, al parecer la temperatura relativa provocó en la edición de 2007 los corredores estuvieran más tiempo esperando en los puntos de control, que en la edición de 2009 que hay una parte de la carrera que se desarrolló con precipitaciones.

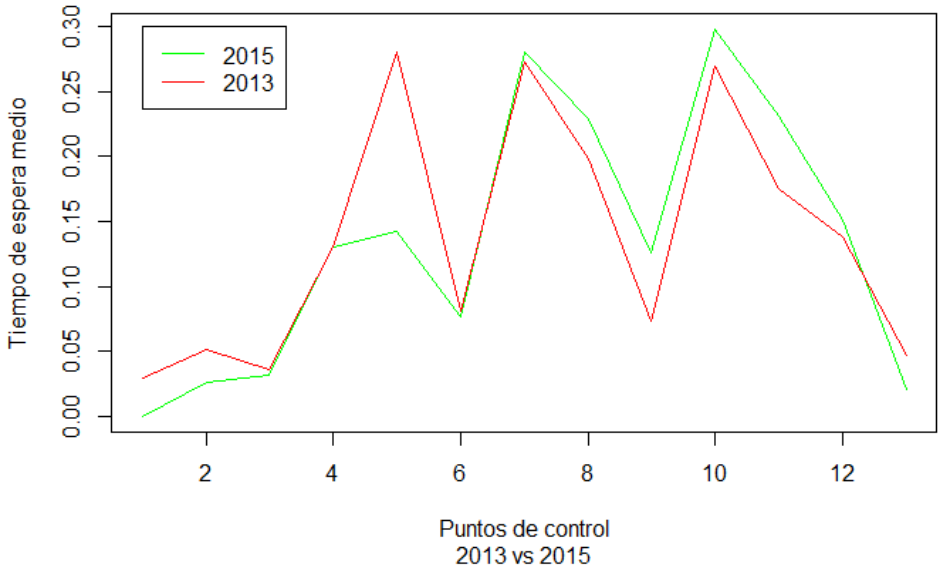


2013-2015

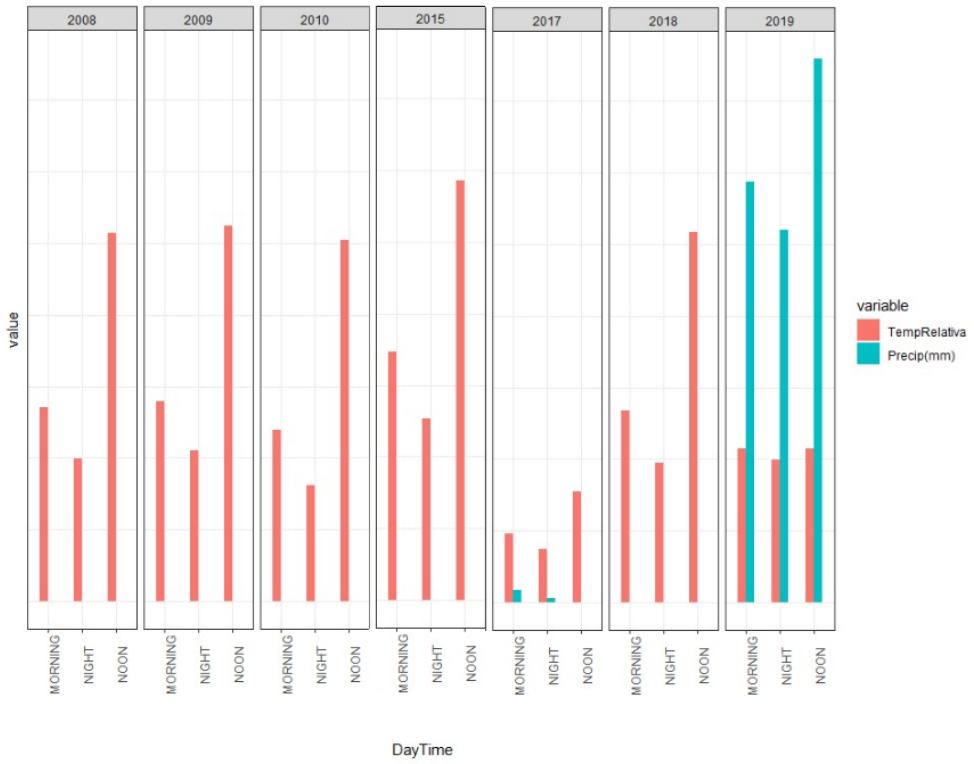
Viendo los años 2013 y 2015, donde en 2013 hubieron precipitaciones (en el rango del medio día) y en 2015 temperaturas bajas, se puede observar como las precipitaciones afectó a las velocidades medias de los corredores de la edición de 2003 (primeros puntos de control,, mientras que en el momento que desaparecieron estas precipitaciones, los datos entre ambas ediciones se vuelven mucho más similares.



En cambio viendo los tiempos de espera entre puntos de control se observa que la edición de 2013 en los puntos que se conoce que hubo precipitaciones sus corredores estuvieron más tiempo esperando en los puntos de control (descansando). Esta dinámica cambia cuando deja de haber precipitaciones en la edición de 2013, la cual a partir de ese punto se asemeja muchísimo a la edición de 2015.



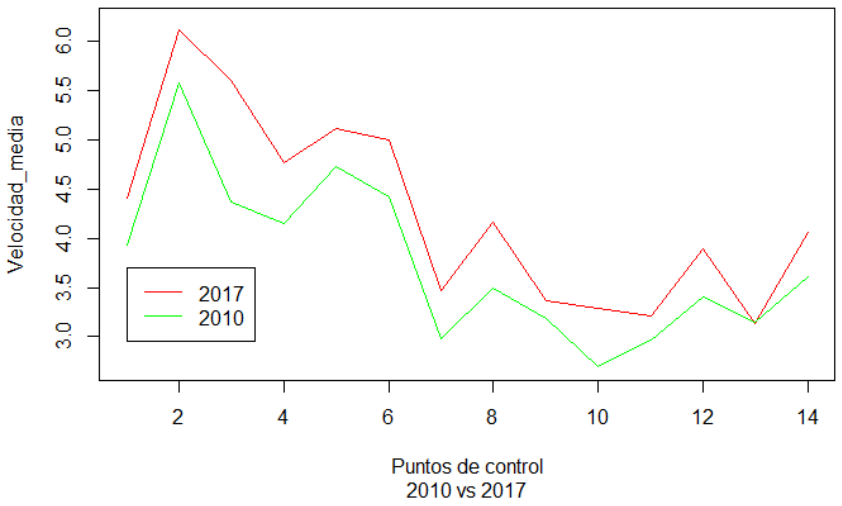
THE BEAR:



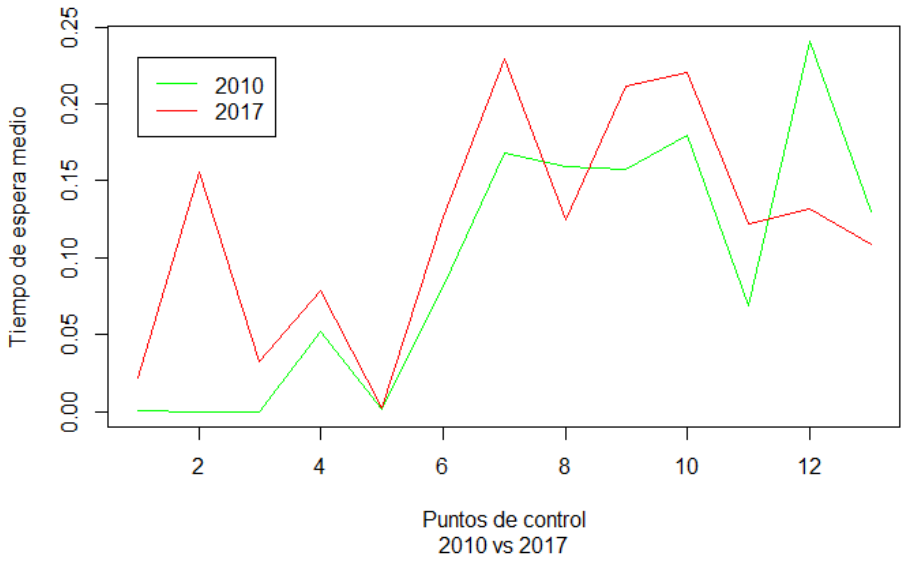
Al ver los gráficos de precipitaciones y de temperaturas relativas, se quiere estudiar cómo afecta las condiciones meteorológicas a los 50 mejores corredores de Bear de cada edición.

2010-2017

Primeramente se comparan los años de 2010 y de 2017 (donde las temperaturas de 2017 son las más reducidas de las diferentes ediciones). Existe una gran diferencia entre la velocidad media de los corredores de la edición del 2017 con la edición del 2010.

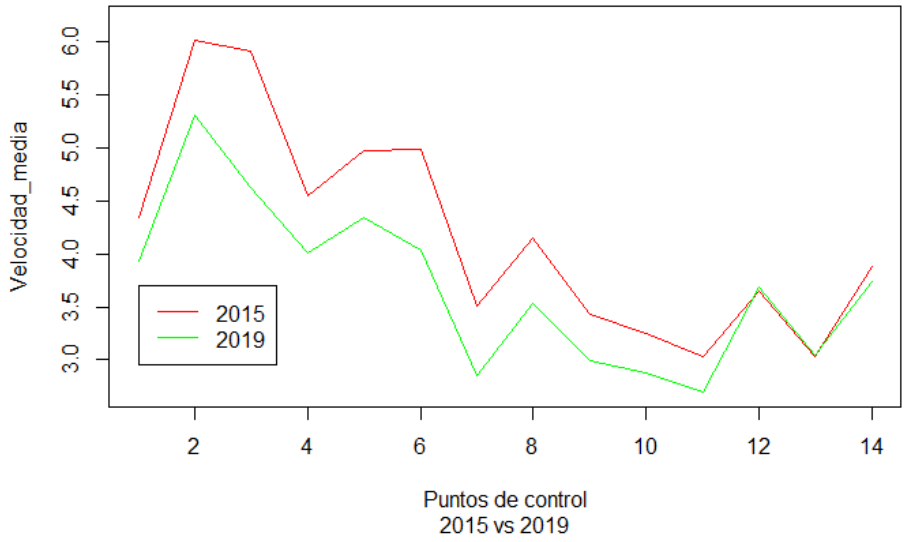


Siguiendo el patrón comentado anteriormente, los tiempos de espera de la edición que menos calurosa (2017) se reduce significativamente respecto a la edición de 2010.

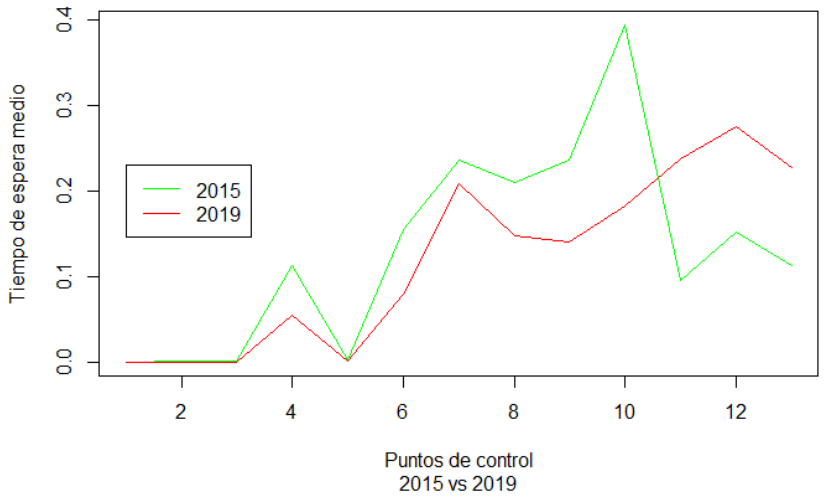


2015- 2019

Las ediciones de 2015 y 2019 se puede ver un cambio remarcable entre la velocidad media de sus participantes, eso puede ser debido ya que en la edición de 2019 han habido una gran cantidad de precipitaciones.

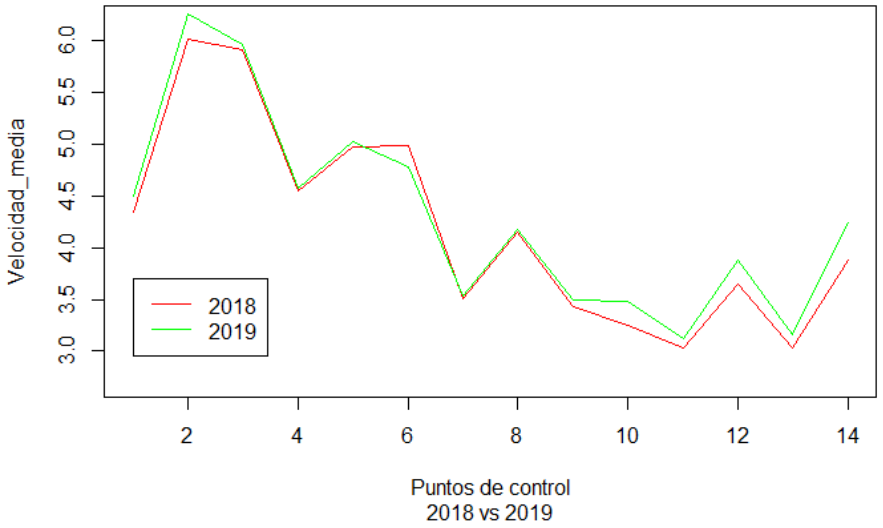


En cambio los tiempos de espera, a pesar de las precipitaciones en la edición de 2019, sus participantes estuvieron menos tiempo descansando en los puntos de control.

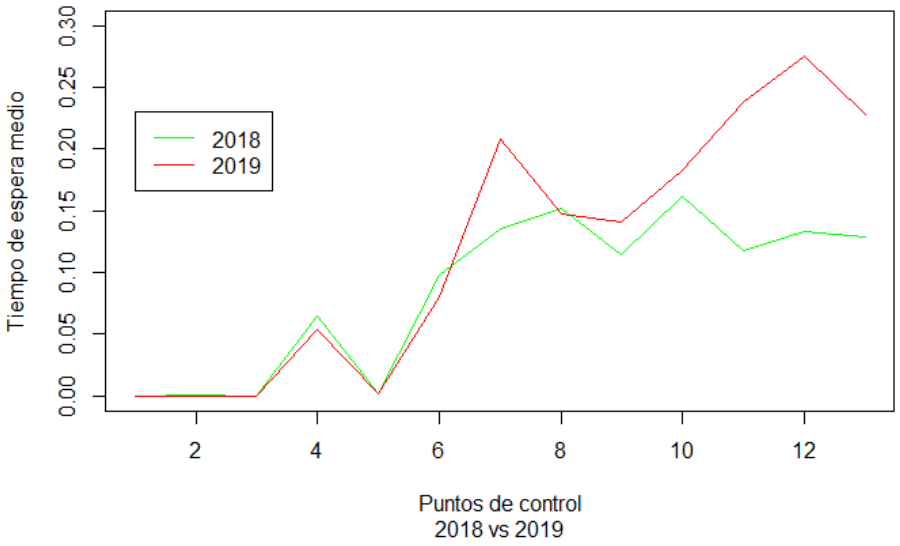


2018-2019

Comparando las ediciones de 2018 y 2019, las cuales respectivamente en una tenía unas condiciones climáticas favorables y la otro tenía una gran cantidad de precipitaciones, como se puede ver en el gráfico las velocidades medias de sus participantes no se vieron muy variadas de una edición con la otra.



En cambio cuando se representan los tiempos de espera, se puede ver una gran diferencia, la edición de 2019 (con precipitaciones) los corredores de media descansaron más tiempo que la edición de 2018 (con condiciones climáticas más favorables).

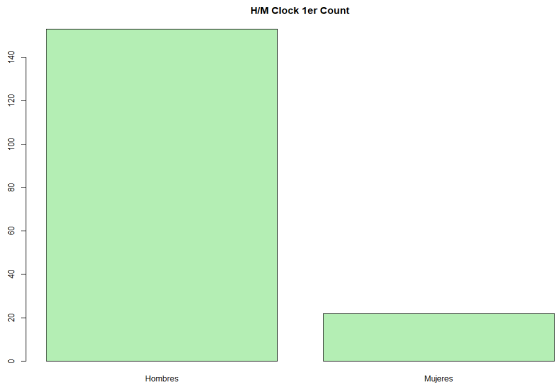


Finalmente se ha de comentar que han habido algunos gráficos que se han decidido obviar ya que su resultado no aportan ninguna información crucial para poder responder a la cuestión que se había planteado.

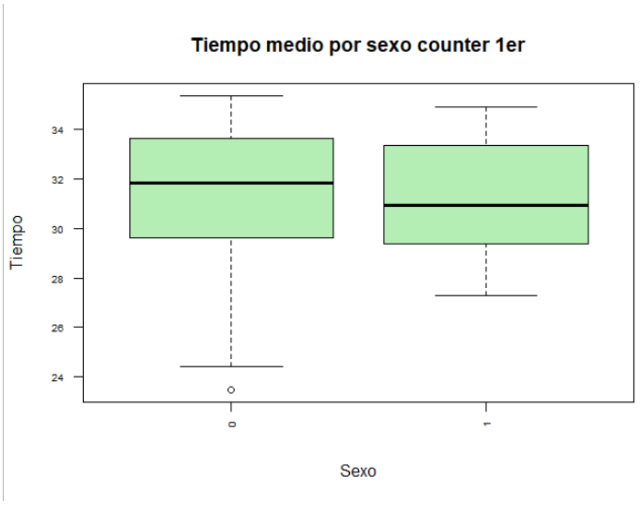
APARTADO 1.2

EFECTO DE EDAD/GÉNERO en Rendimiento

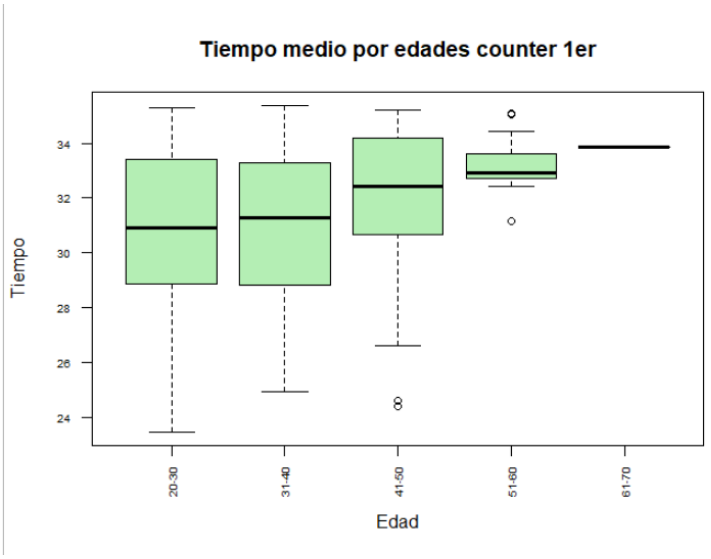
1. ***1.2.1***



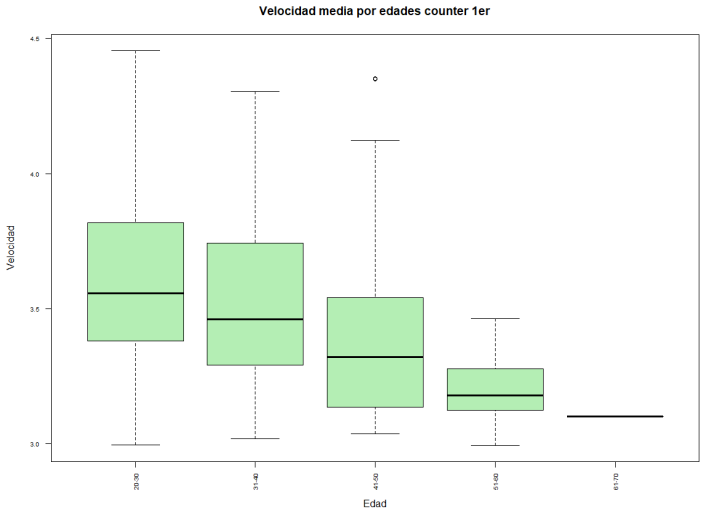
1. ***1.2.2***



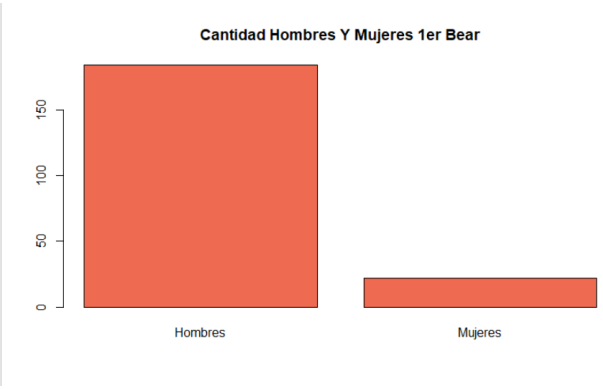
1. ***1.2.3***



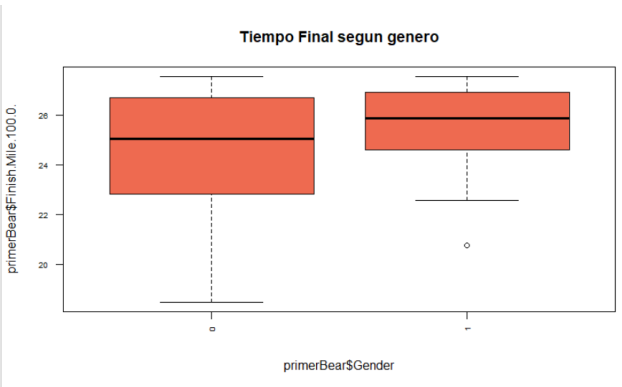
1. **1.2.4**



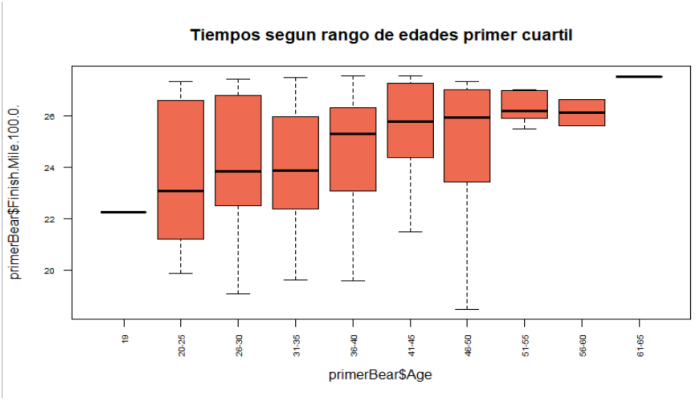
1. **1.3.1**



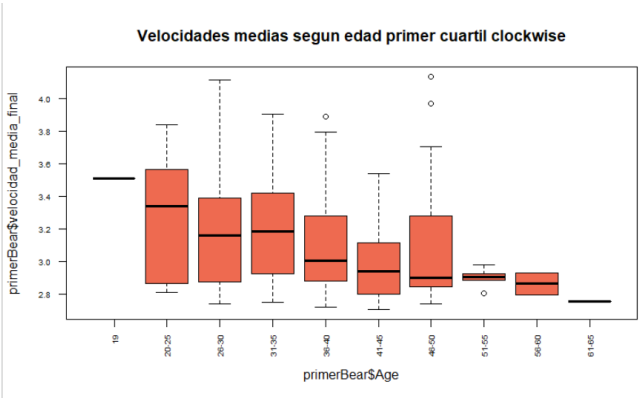
1. **1.3.2**



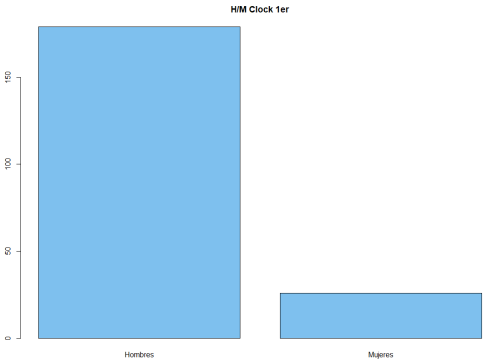
1. **1.3.3**



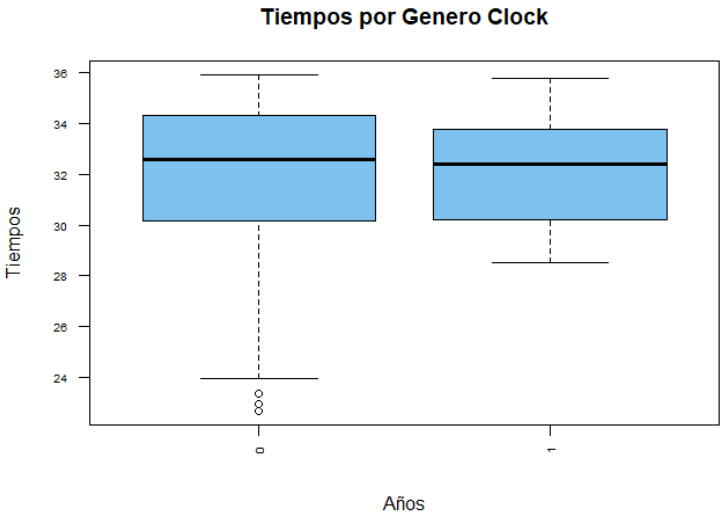
1. **1.3.4**



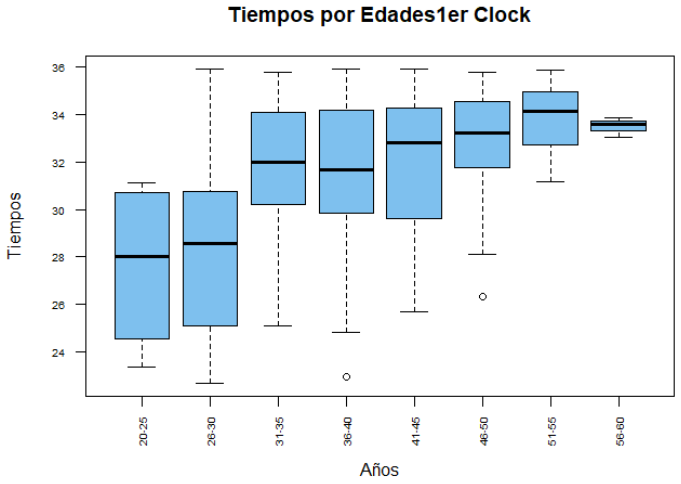
1. **1.3.5**



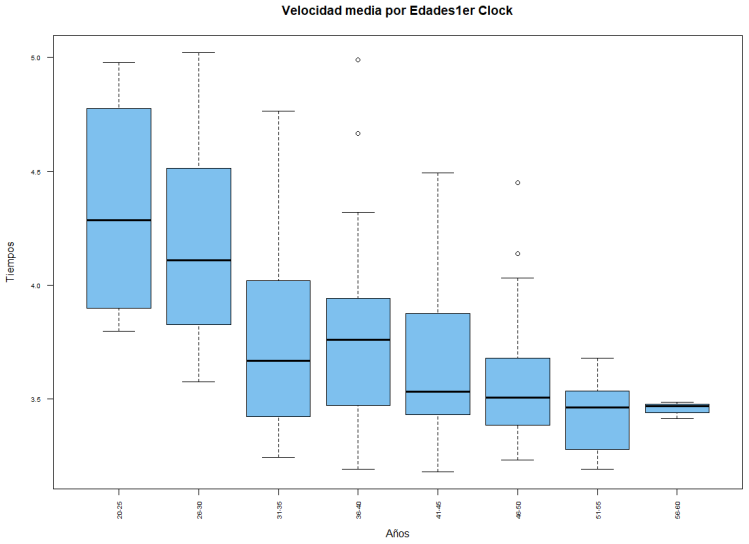
1. **1.3.6**



1. **1.3.7**



1. **1.3.8**



APARTADO 1.3

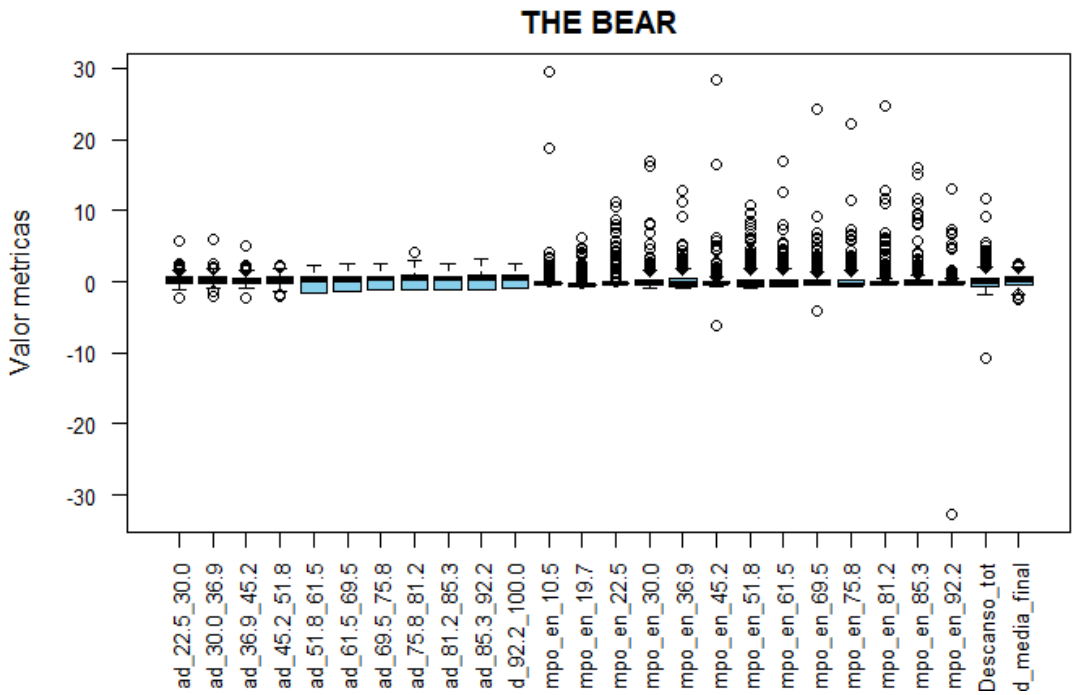
PCA

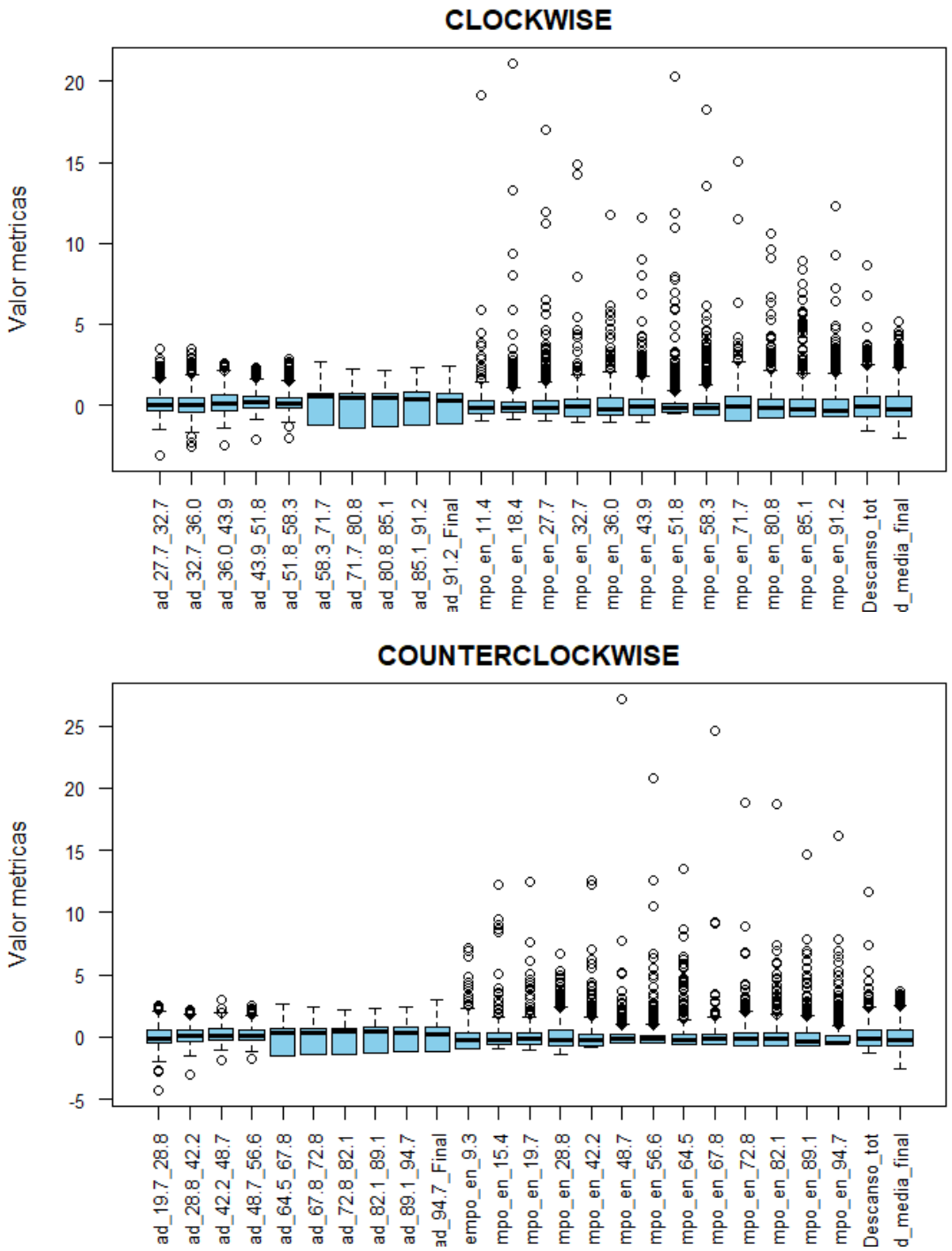
Para estudiar los resultados, y el efecto de las variables en la explicación de la variabilidad de los datos, vamos a realizar un análisis de componentes principales.

Vamos a eliminar los tiempos de entrada y salida en cada punto de control por ser variables temporales aditivas que no podremos estudiar. sin embargo vamos a realizar el PCA, con las velocidades medias por etapa, y los tiempos de descanso en cada punto de control, para determinar si su valor en ciertos puntos a lo largo de la carrera, son más determinantes.

Como vamos a intentar estudiar la variabilidad de los datos, y dado que los datos están medidos en distintas métricas(velocidad media por milla/ tiempo en minutos/horas) esto puede alterar el análisis, ya que se le puede dar más importancia a algunas variables por el hecho de tener un rango más amplio, lo que es posible que distorsione nuestros resultados.

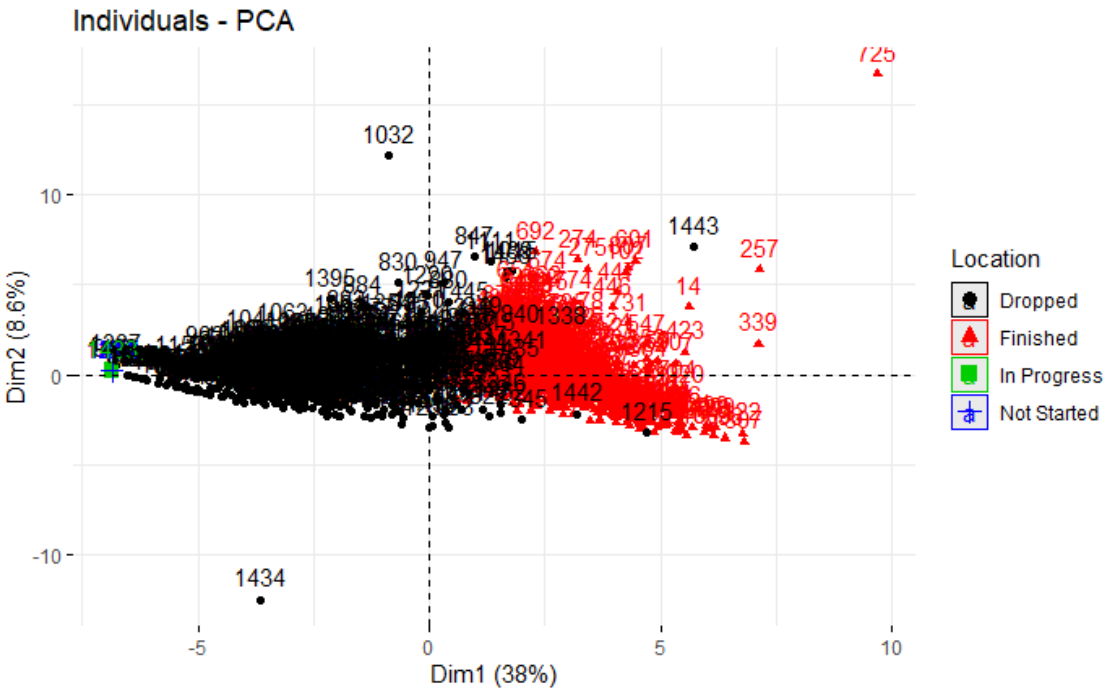
Para evitar que esto pase, vamos a escalar y centrar los datos de los que disponemos, para que la media de cada variable sea 0, y su desviación estándar 1.





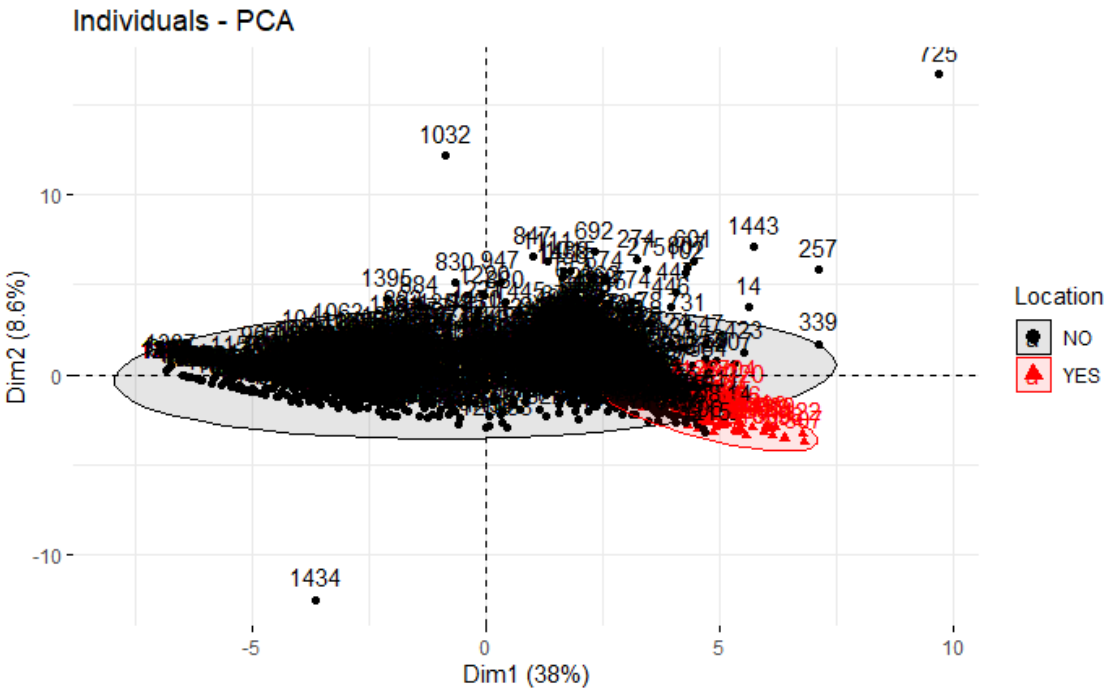
Tras esto, calculamos sus proyecciones en el nuevo espacio formado por las dos primeras componentes, y vamos a estudiar cómo se distribuyen los scores para los corredores que terminan la carrera/abandonan, o están en el top 10 de su año, para luego determinar qué variables determinan esto.

THE BEAR

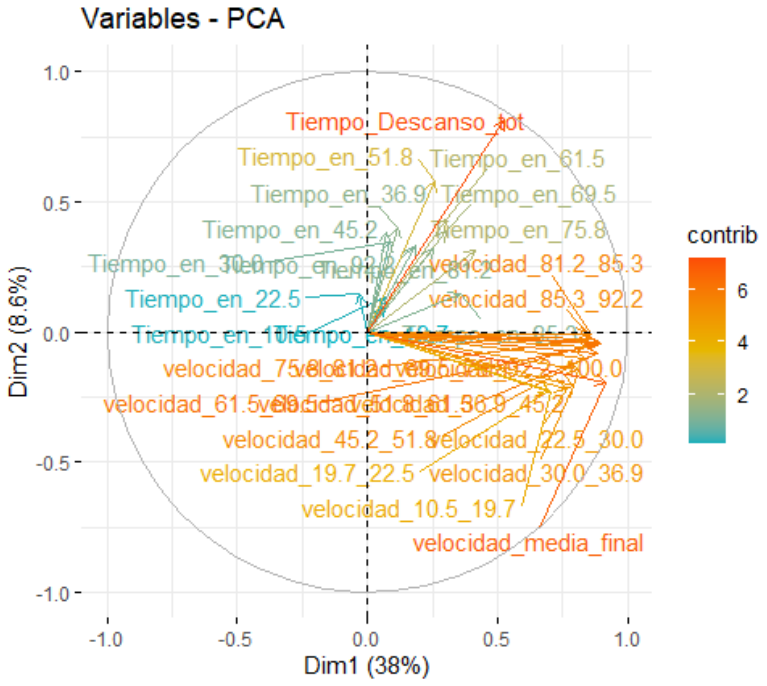


Visualizamos los scores para los corredores coloreando según su estado al final de la carrera para el circuito The Bear, podemos ver que los datos de los corredores se distribuyen a lo largo del eje de la primera componente, por lo que su variabilidad está explicada por esta dimensión, algo de variabilidad también puede ser explicada por la segunda componente, aunque menos, ya que vemos algunos scores dispersandose según el eje Y.

Vemos que la dimensión 1 explica un 38% de la variabilidad, mientras que la dimensión 2 solamente un 8.6%, en cuanto a los datos de los corredores que han terminado en comparación con los corredores que abandonan,la variabilidad de los que logran terminar la carrera está más explicada por la segunda componente, ya que su dispersión por dicho eje es un poco mayor. A la hora de distinguir ambos grupos, existe algo de solapamiento, y algún anómalo como el score 1443, pero en general los scores de los corredores que terminan, son más altos en la primera componente, como se posicionan más a la derecha en la primera dimensión, asumimos que las variables que más contribuyen a su explicación son las que separan a dichos corredores.



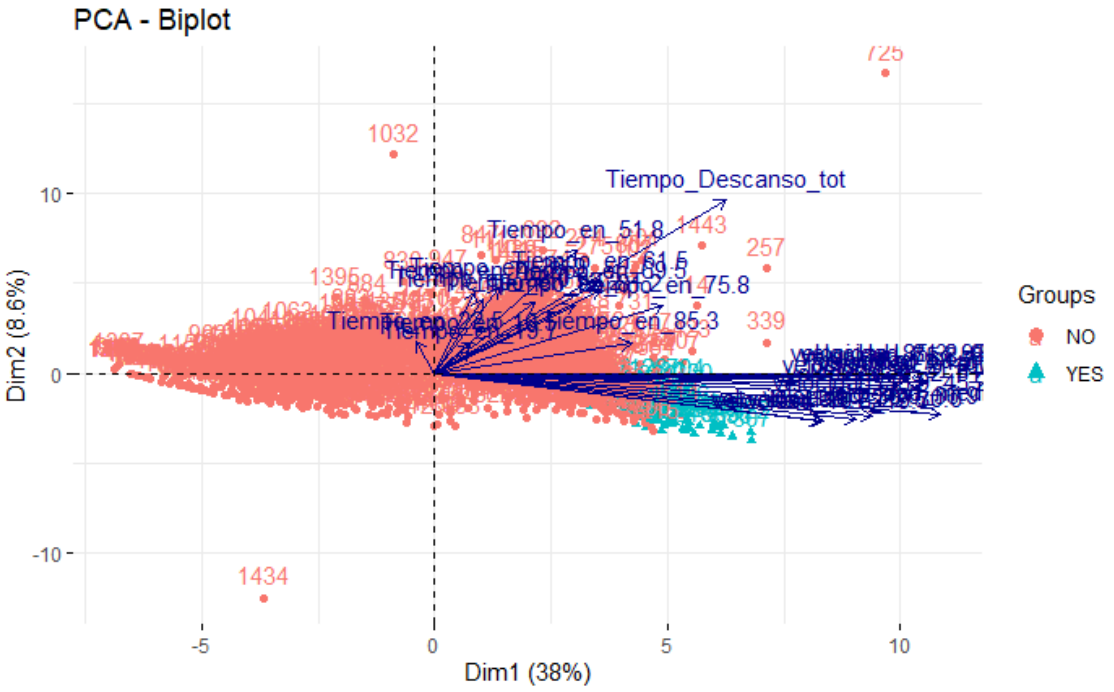
Visualizamos ahora los 10 mejores corredores de cada carrera para cada año del evento, y vemos que los scores de dicho grupo se posiciona lo más alto posible a lo largo de la primera componente, y su variabilidad se ve explicada mayoritariamente por la primera dimensión ya que ya que se distribuye a lo largo del eje x, pero también se ve más afectada por la segunda dimensión, por lo que las variables que influyen en dichas componentes, son aquellas que diferencian los mejores del resto de corredores. vamos a ver ahora qué variables son estas.



Cabe indicar que el gráfico proporcionado por el paquete FactoMineR no muestra los loadings de la variables, sino los loadings multiplicados por la desviación típica de la componente correspondiente, es decir, la raíz cuadrada del valor propio asociado. De esta forma, los valores representados en el gráfico (coordenadas) están penalizados por el porcentaje de variable

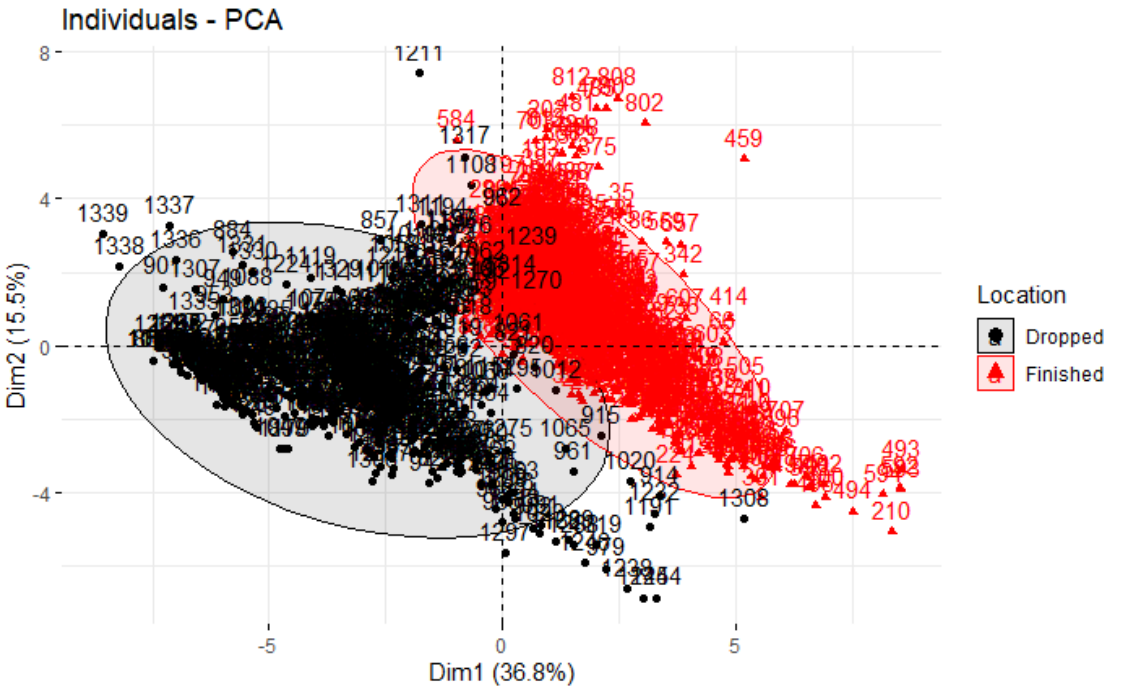
explicada por la correspondiente PC y, dado que las variables están tipificadas, representan la correlación de cada una de las variables originales con las PCs.

Vemos que las velocidades en los puntos de control son las que más influyen en la primera dimensión(que explica 38% de la variabilidad de los datos para los corredores), siendo las velocidades para los puntos de control al final de la carrera, entre las millas 81.2-85.3, 85.3-92.2, 92.2-100,las más pegadas al eje, por los que están más correlacionadas con esta componente y como tienen bastante contribución, las que más peso tienen en esta componente, claro está exceptuando la velocidad media final, que está menos correlacionada con la dimensión 1 pero su contribución es mayor.
Por lo que estas serían las que más diferenciarán entre los que acaban o no la carrera, y los top 10 del resto, también contribuyen a la explicación de la primera componente, las velocidades al inicio de la carrera, pero están menos correlacionadas con esta dimensión, al tener un ángulo mayor con el eje.
Por otro lado, la variabilidad en la segunda dimensión (8.6% de la variabilidad de los datos)viene determinada por los tiempos de descanso de los corredores, pero vemos que tienen poca contribución a la dimensión 2, exceptuando el tiempo de descanso total, y el tiempo de descanso en la milla 51.8, pero estas guardan poca correlación con esta dimensión por no estar muy pegadas a su eje.

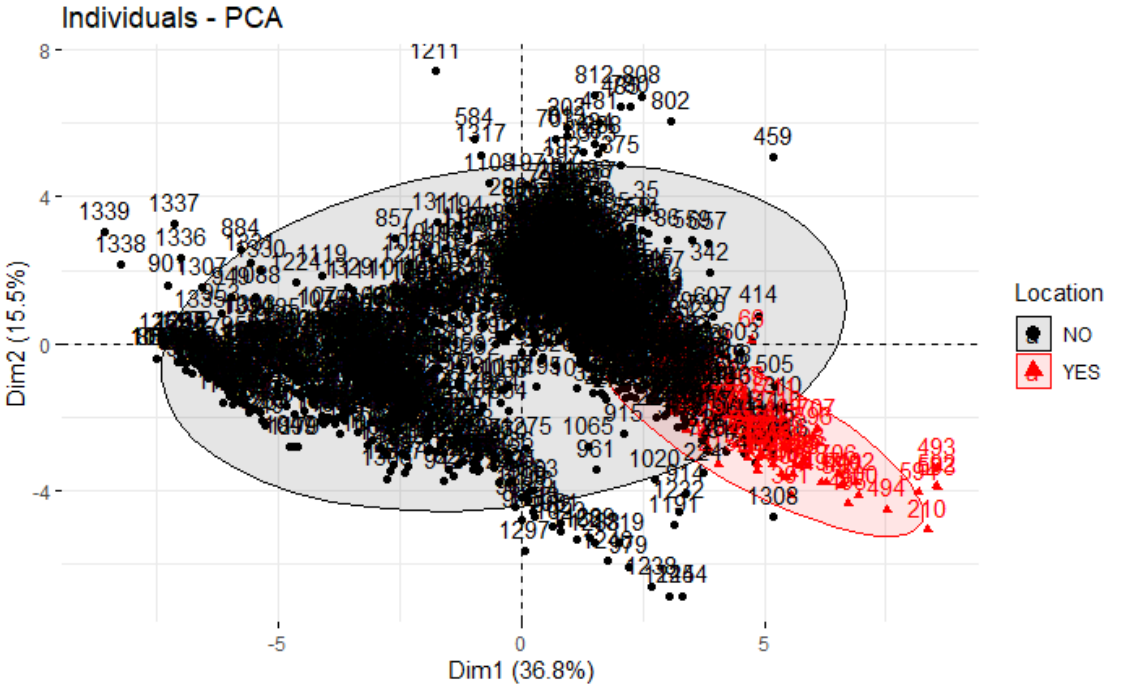


En este último gráfico, en el cual coloreamos en si el corredor es top 10 de su año y ahora si, los loadings de las variables, vemos cómo los tiempos de descanso, y sobre todo el tiempo de descanso de un corredor, es aquel que distribuye los datos hacia la esquina derecha superior, mientras que las velocidades medias entre las etapas diferencian a los corredores top, e influyen en su distinción en la primera componente hacia la derecha.

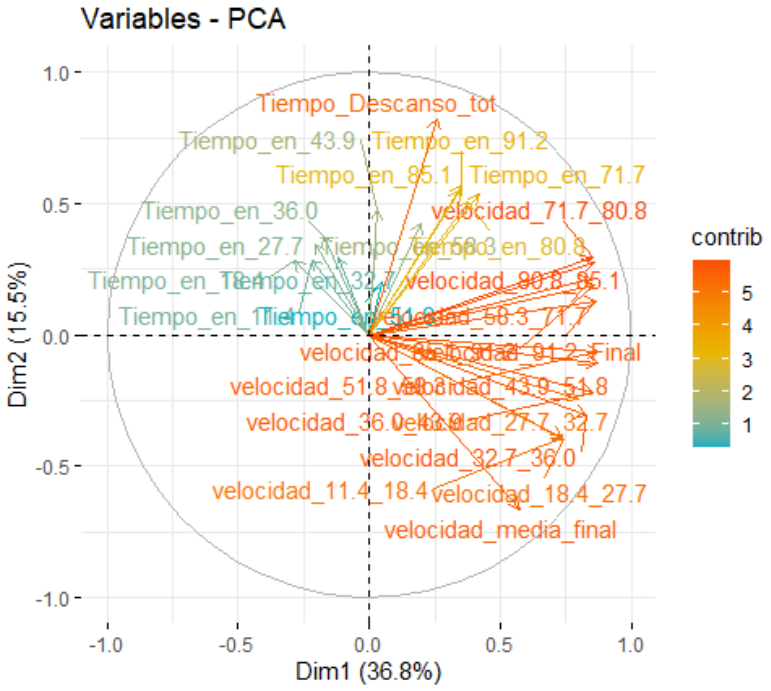
CLOCKWISE HARD ROCK



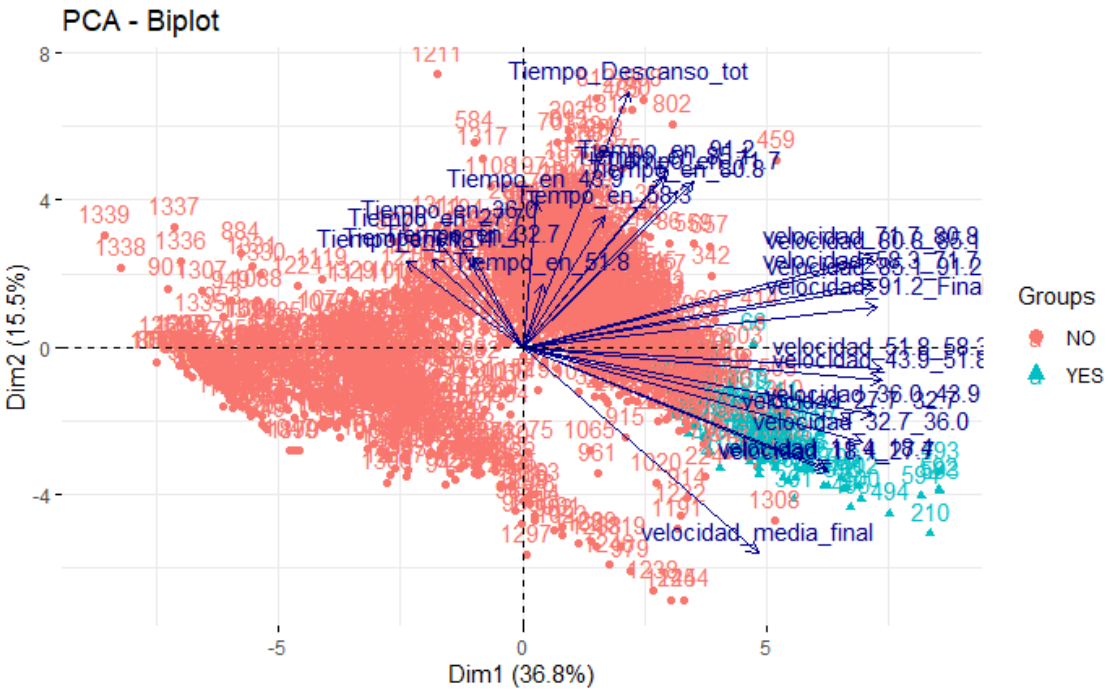
Vemos cómo se distribuyen los scores según el estado final del corredor, y vemos que para esta carrera, da variabilidad de los scores viene determinada por ambas dimensiones, la dimensión 1 explica el 36.8% de la variabilidad de los datos, mientras que la dimensión 2 explica un 15.5%, más que en la carrera The Bear. Vemos que los que acaban la carrera y los que no, están diferenciados sobre todo por la primera componente, aunque podemos ver una mayor distribución por el eje de la segunda dimensión para los scores de los corredores que la terminan, pero en general las variables que explican la primera componente, son las que permiten distinguir los corredores que terminan, y los que no.



Vemos en el gráfico coloreado según los 10 mejores corredores cada año, que al igual que en the bear, los mejores están distribuidos más a la derecha que los demás, con scores más altos para la primera componente, esto es debido a la influencia de las variables que influyen en esta componente principal. Vamos ahora a estudiar las variables que explican las dimensiones.



En comparación con la carrera The Bear, vemos que para este evento, los tiempos de descanso tienen más peso en la segunda componente, como el tiempo de descanso total, el tiempo de descanso en las millas 85, 91.2, 71.7 y 80.8, pero exceptuando el tiempo de descanso en la milla 43.9, que si se correlaciona con esta dimensión, pero no contribuye mucho a su explicación, el resto de variables de tiempo de descanso, no están completamente correlacionadas. El tiempo de descanso total en esta carrera está bastante más correlacionado con la componente 2 que en la carrera The Bear, esto puede deberse a que es un circuito más duro, por lo que es más importante descansar para tener mejor rendimiento.
En cuanto a las velocidades medias observamos que al igual que en bear, están proyectadas en la dirección de la primera componente, por lo que explican dicha dimensión, a pesar de que en esta carrera, no como en the bear, no hay ninguna proyección pegada justo en el eje. vemos que la velocidad media que más correlacionada está con esta componente es la velocidad desde la milla 91.2 hasta el final de la meta, en cuanto a la contribución de las velocidades, son mayores para las etapas entre las millas 71.8-80.8, 90.8-95.1, que son las más cercanas al final de la carrera, igual que en The Bear, podemos ver una progresión, cuanto más cerca del inicio de la carrera está una etapa, menos correlacionada con la primera dimensión, y menos contribución tiene.



Observando ahora los loadings, vemos cómo a pesar de que las velocidades medias de las etapas finales están más correlacionadas con la dimensión 1, vemos que los corredores to se distinguen del resto de corredores, según sus resultados en las velocidades medias en las primeras, y medias etapas (32.7-36, 18.4-27.7 ...etc) y que su variabilidad a lo largo de este eje viene determinado por estas, un mejor resultado en ellas, diferencia entre los corredores top y el resto, vemos también como la variabilidad a lo largo del eje de la segunda dimensión, de los corredores que terminan la carrera, viene determinada por el tiempo de descanso total, y los tiempos de descanso de las etapas, 91.2, 50.8 58.7… etc.
para finalizar es curioso ver como la velocidad media final está negativamente correlacionada con los tiempos de descanso en algunas etapas iniciales de la carrera ya que las proyecciones de sus pesos están en un ángulo de 180, por lo que deducimos que cuanto más descanso en las etapas iniciales implica una menor velocidad media, ya que el corredor tiene que poder llegar a un ritmo estable para poder tener un buen rendimiento a lo largo de la carrera.

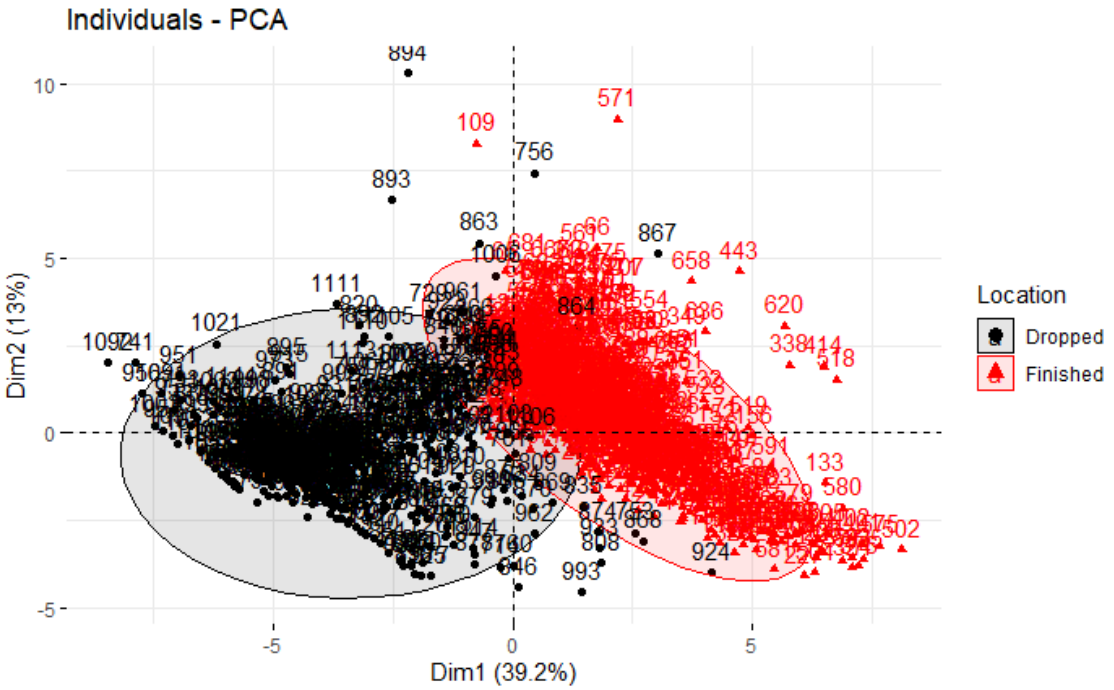
COUNTERCLOCKWISE HARD ROCK

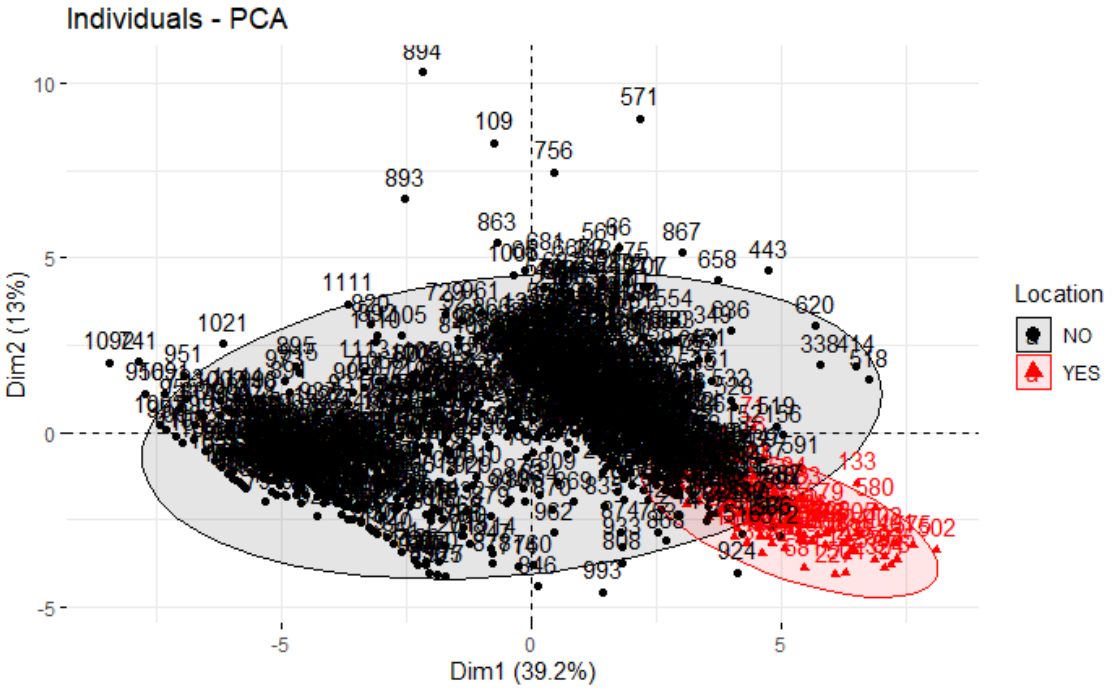
Estudiando los siguientes gráficos, podemos observar los patrones similares entre Hard Rock Clockwise y Hard Rock Counter Clockwise, al solo diferenciarse el recorrido en algunas etapas, a pesar de esto hay algunas diferencias significativas que vamos a señalar.

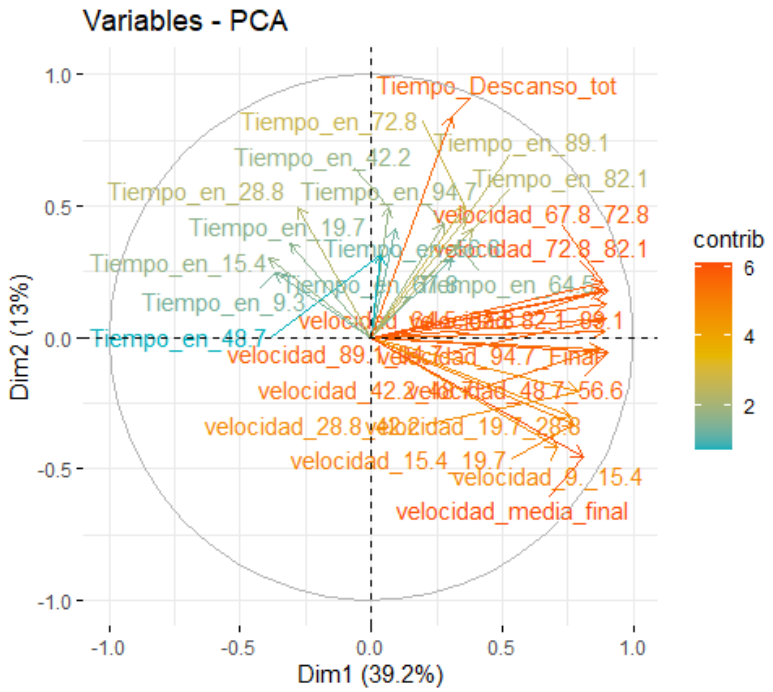
En general, observamos una menor distribución de la variabilidad, a lo largo del eje de la dimensión 2, y una mayor presencia de scores con valores elevados para la primera componente. También podemos ver una mayor distinción de los corredores top de la carrera con el resto, al haber un mayor número de scores fuera de la elipse de los corredores normales, por lo que hay menos scores solapados en ambos grupos.

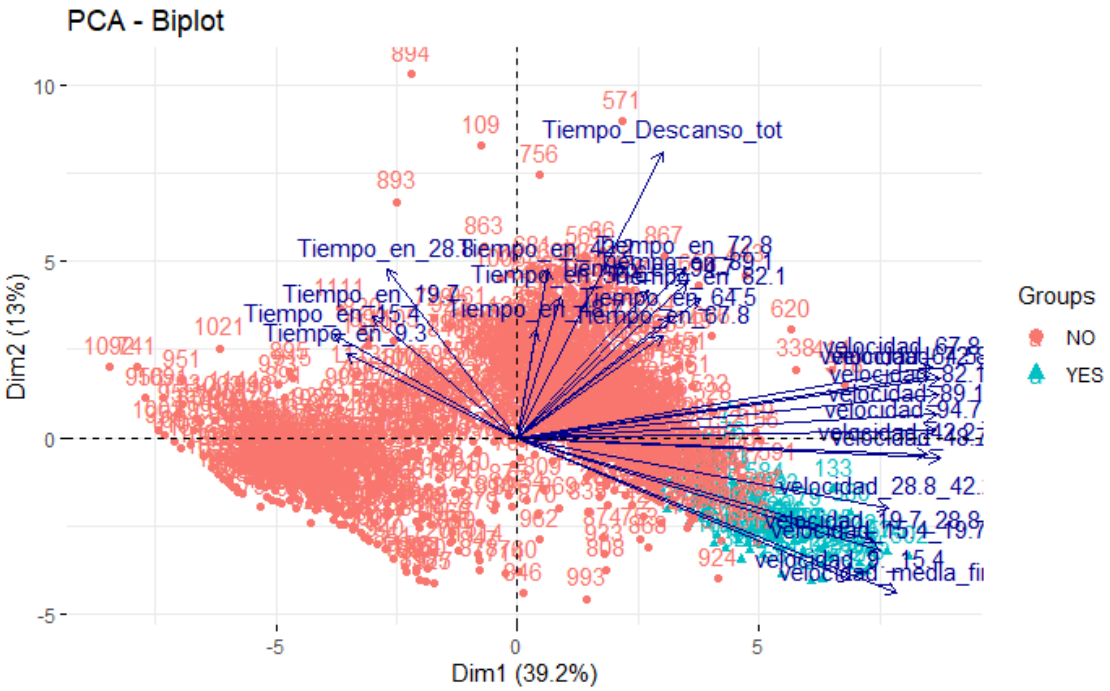
En cuanto a las variables que explican las componentes, los tiempos de descanso en las etapas,tienen una menor contribución en comparación con Clockwise a la dimensión 2, aunque el tiempo de descanso total sigue estando algo correlacionado y contribuye a la explicación de esta componente.

En la primera componente, vemos como las velocidades medias para las etapas iniciales separan a los corredores top del resto, pero además ahora la velocidad media final está más correlacionada con la componente principal.









APARTADO 1.4

Analisis PLS -DA

Deseamos saber si con los parámetros recopilados que describen la actuación de un corredor durante ciertas etapas de la carrera, podemos predecir el estado final del corredor (dropped/ finished) , para ello vamos a realizar un análisis discriminante PLS, para entrenar un modelo y ver si es posible con este predecir correctamente si un corredor va a abandonar o no una carrera.

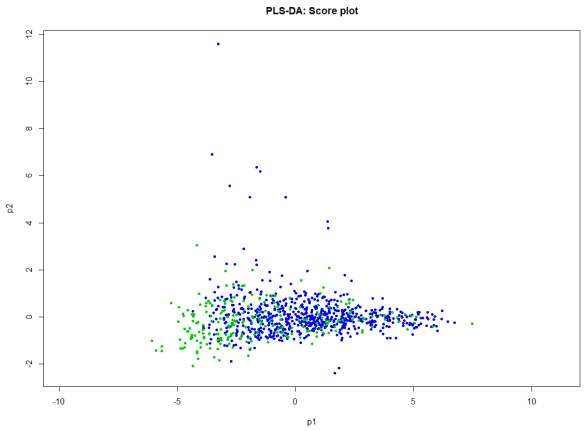
decidimos coger las medidas para los puntos de control desde aproximadamente la milla 30, hasta la 50/60, ya que para las primeras millas, los corredores aún no han sido puestos a prueba y solo los lesionados, o los primerizos que no daban la talla han abandonado la carrera, procedemos pues a partir nuestros datos en dos sets, uno siendo el 80%, que usaremos para entrenar el modelo, y otro siendo el 20% que usaremos para aplicar y comprobar su efectividad.

Bear:

RMSEE is the square root of the mean error between the actual and the predicted responses

```
880 samples x 12 variables and 1 response
standard scaling of predictors and response(s)
      R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
Total    0.654    0.125   0.112 0.409   2   0 0.05 0.05
> myplsda@modelDF
     R2X R2X(cum)    R2Y R2Y(cum)      Q2 Q2(cum) Signif. Iter.
p1 0.565    0.565 0.1080    0.108 0.10400   0.104      R1     1
p2 0.089    0.654 0.0177    0.125 0.00861   0.112      R1     1
```

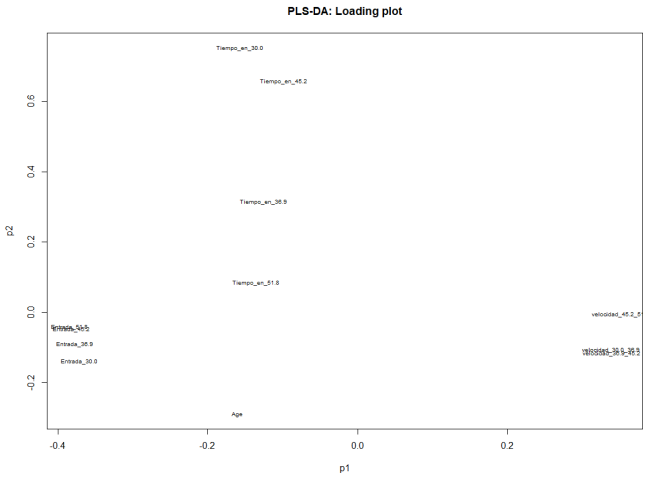
podemos ver que para Bear, nos crea un modelo con dos componentes principales explicando su variabilidad y poca capacidad predictiva, ya que Q2 es de 0.112, y con un error medio de 0.409 entre el desenlace actual y el predecido lo que nos indica la poca fiabilidad del modelo creado, sin embargo la cantidad de variabilidad que explican nuestros datos es elevada, pero solamente para la dimensión 1 ya que R2X = 0.654 Y R2Y = 0.125, entonces las variables que influyen en dicha componente principal, son las que más explican la variabilidad en nuestros datos, podemos visualizar esto en el gráfico de scores.



Vemos como los puntos se distribuyen mayoritariamente a lo largo del eje X (primera dimensión) a pesar de que hay algo de distribución por el eje Y también, los scores verdes son aquellos corredores que han abandonado, y los azules, los que han logrado acabar la carrera, una vez visto esto, vamos a ver qué variables son las que influyen en las dimensiones.

Aquí vemos como las velocidades entre las millas 30 - 36.9, 36.9 - 45 y 45.2 - 51.8 tienen mucho peso en la explicación de la primera componente principal, como esta es la que diferencia los estados de los corredores, las velocidades en estas etapas pueden diferenciar si el corredor acaba o no acaba la carrera, en cuanto a la variabilidad de los scores a lo largo de la segunda dimensión, los tiempos de descanso en la milla 30 y 45.2 son las responsables.

Probamos el modelo en los datos test, donde obtenemos estos resultados.




```
Confusion Matrix and Statistics

          mypred2
Y2         Dropped Finished
  Dropped        9       47
  Finished       5      158

               Accuracy : 0.7626
                 95% CI : (0.7006, 0.8173)
    No Information Rate : 0.9361
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1725

 Mcnemar's Test P-Value : 1.303e-08

            Sensitivity : 0.64286
            Specificity : 0.77073
         Pos Pred Value : 0.16071
         Neg Pred Value : 0.96933
             Prevalence : 0.06393
         Detection Rate : 0.04110
   Detection Prevalence : 0.25571
      Balanced Accuracy : 0.70679

       'Positive' Class : Dropped
```

Con los 219 corredores de los datos test, el modelo acierta sus predicciones un 76% de las veces, El índice Kappa, que mide cómo excede el modelo en términos de exactitud a predicciones aleatorias es de 0.17 lo cual indica unas predicciones no muy alejadas de la aleatoriedad, por lo que con los datos extraído de esas etapas no son suficientes para predecir exactamente si un corredor va a finalizar o abandonar una carrera.

Clockwise:

```
894 samples x 12 variables and 1 response
standard scaling of predictors and response(s)
      R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
Total    0.724    0.173    0.14 0.404   3   0 0.05 0.05
> myplsda@modelDF
      R2X R2X(cum)    R2Y R2Y(cum)      Q2 Q2(cum) Signif. Iter.
p1 0.6080    0.608 0.1040    0.104 0.10000   0.100      R1     1
p2 0.0685    0.677 0.0482    0.152 0.04090   0.137      R1     1
p3 0.0471    0.724 0.0205    0.173 0.00342   0.140      R1     1
```

```
          mypred2
Y2         Dropped Finished
  Dropped       16       44
  Finished      27      136

               Accuracy : 0.6816
                 95% CI : (0.6161, 0.7422)
    No Information Rate : 0.8072
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.111

 Mcnemar's Test P-Value : 0.05758

            Sensitivity : 0.37209
            Specificity : 0.75556
         Pos Pred Value : 0.26667
         Neg Pred Value : 0.83436
             Prevalence : 0.19283
         Detection Rate : 0.07175
   Detection Prevalence : 0.26906
      Balanced Accuracy : 0.56382

       'Positive' Class : Dropped
```

CounterClockwise:

```
PLS-DA
981 samples x 16 variables and 1 response
standard scaling of predictors and response(s)
      R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
Total    0.688    0.167   0.137 0.431   2   0 0.05 0.05
```

```
Confusion Matrix and Statistics

          mypred2
Y2         Dropped Finished
  Dropped       20       61
  Finished      26      137

               Accuracy : 0.6434
                 95% CI : (0.5798, 0.7035)
    No Information Rate : 0.8115
    P-Value [Acc > NIR] : 1.0000000

                  Kappa : 0.0981

 Mcnemar's Test P-Value : 0.0002672

            Sensitivity : 0.43478
            Specificity : 0.69192
         Pos Pred Value : 0.24691
         Neg Pred Value : 0.84049
             Prevalence : 0.18852
         Detection Rate : 0.08197
   Detection Prevalence : 0.33197
      Balanced Accuracy : 0.56335

       'Positive' Class : Dropped
```

APARTADO 1.5

Como podemos ver en el gráfico anterior, no existe ninguna correlación significativa excepto la velocidad media y el tiempo final , esto era de esperar ya que a mayor velocidad mayor velocidad media por lo tanto menor tiempo final , el resto de correlaciones no tienen un alto nivel de significación,solamente destacar que el tiempo final y el descanso tienen una relación aunque es muy débil.

BIBLIOGRAFÍA:

Páginas web que ha juicio de los integrantes del estudio han sido más importantes.

Artículo de cómo afecta la temperatura a los corredores de montaña

https://www.carreraspormontana.com/salud/como-afecta-el-calor-al-corredor-de-montana-segun-el-rango-de-temperatura/

Artículo sobre entrenamiento deportivo

https://www.sportlife.es/mujer/entrenamiento-mujer/articulo/entrenamiento-edad-claves-alargar-vida-deportiva

Página web sobre carreras de montaña

https://carrerasdemontana.com/

Página web con muchas carreras de montaña (como se ha dicho en el estudios de esta página de donde se han podido recabar los datos utilizados en el estudio)

https://www.opensplittime.org/

Manual Python

https://www.geodose.com/2018/04/create-gpx-tracking-file-visualizer-python.html

Otras páginas web utilizadas.

https://www.alltrails.com/es/explore/recording/hardrock-100-gpx?u=m

https://www.alltrails.com/es/explore/trail/us/utah/the-bear-100-ultra?u=m

https://chronorace.tracktherace.com/es/transgrancanaria-2019-360/race

https://www.ultratrail-worldtour.com/es/corredores/clasificacion-ultra-trail-world-tour/clasificacion-anual-2017/

https://www.trailforks.com/

https://runealo.es/se-cancela-la-mitica-hardrock-100/

https://runealo.es/se-cancela-la-mitica-hardrock-100/