ARMONIZACIÓN DE LAS ESTADÍSTICAS PROVINCIALES DE TURISMO

DOCUMENTO TÉCNICO #6

Ciencia de Datos para el Turismo

Dirección Nacional de Mercados y Estadística
SUBSECRETARÍA DE DESARROLLO ESTRATÉGICO

Ministerio de Turismo y Deportes
Argentina

Documento Técnico N°6: Ciencia de Datos para el Turismo

El turismo desde la perspectiva de la oferta: actividad y empleo

Dirección Nacional de Mercados y Estadísticas - Subsecretaría de Desarrollo Estratégico

29 de septiembre de 2021

Índice general

Presentación

El presente documento, **Ciencia de Datos para Turismo**, se enmarca en el proyecto de Armonización de las Estadísticas de Turismo en las Provincias de la Dirección Nacional de Mercados y Estadística de la Subsecretaría de Desarrollo Estratégico del Ministerio de Turismo y Deportes. El objetivo general de este proyecto es contribuir con propuestas metodológicas para los sistemas de estadísticas de turismo provinciales que orienten a producir indicadores provinciales básicos y comparables.

Además de este, se encuentra disponible una serie de documentos técnicos que abordan otras problemáticas vinculadas a la producción de estadística de turismo:

- Documento Técnico #1: Conceptos y elementos básicos para la medición provincial de los turistas
- Documento Técnico #2: Propuestas metodológicas para las encuestas de ocupación en alojamientos turísticos
- Documento Técnico #3: Descripción, análisis y utilización de los Registros Administrativos para la medición del Turismo
- Documento Técnico #4: Propuestas Metodológicas para las Encuestas de Perfil del Visitante
- Documento Técnico #5: Medición de la contribución económica del turismo: actividad y empleo

Documento Técnico Nº6 - Resumen

La ciencia de datos es una disciplina que ha brindado nuevas y maravillosas posibilidades a muchas industrias por medio de la explotación de datos. Junto con estas posibilidades, también ha traído consigo cambios y desafíos constantes. La industria del turismo no es una excepción.

En este documento técnico realizaremos una introducción al concepto de ciencia de datos y su proceso. Introduciremos el lenguaje de programación R como la

caja de herramientas principales para poder llevar adelante cada tarea y etapa de este proceso.

El documento se divide en xx capítulos con ejemplos prácticos y ejercicios (desafíos) para introducir y practicar los conceptos mencionados.

Capítulo 1

Ciencia de Datos

> "Disciplina **emergente** que se basa en el conocimiento en **metodología estadística y ciencias de la computación** para crear predicciones, clasificaciones e ideas impactantes para una amplia gama de campos tradicionales"

No existe un acuerdo sobre una definición formal de ciencia de datos, pero la mayoría de estas definiciones concuerda en que tiene al menos tres pilares: el conocimiento estadístico, el conocimiento de ciencias de la computación y el conocimiento del negocio sobre el cual se va a aplicar. En este caso el turismo.

El proceso de ciencia de datos en el cual nos vamos a basar se puede ver en el siguiente diagrama:

Primero, debes **importar** tus datos hacia la herramienta donde vas a procesarlos. Típicamente, esto implica tomar datos que están guardados en un archivo o base de datos y cargarlos en tu software para poder trabajar con ellos.

Una vez que has importado los datos, el siguiente paso es **ordenarlos** para que tengan un formato adecuado para su análisis. Este formato pensado para el análisis tiene la característica que, en los set de datos ordenados, *cada columna es una variable y cada fila una observación*. Tener datos ordenados nos provee una estructura consistente, preparada para analizarlos y podemos enfocar nuestros esfuerzos en las preguntas que queremos contestar con nuestros datos y no tener que acomodarlos cada vez que la pregunta cambie.

Cuando tus datos están ordenados, podemos necesitar *transformarlos*. La transformación implica quedarte con las observaciones que sean de interés (como todos los hoteles de una ciudad o todos los datos del último año), crear nuevas variables que a partir de variables ya existentes (como calcular el porcentaje de ocupación a partir de la cantidad de plazas totales y las ocupadas) y calcular una serie de estadísticos de resumen (como recuentos y medias).

Una vez que tienes los datos ordenados con las variables que necesitas, hay dos principales fuentes generadoras de conocimiento: la **visualización** y el **modelado**. Ambas tienen fortalezas y debilidades complementarias, por lo que cualquier análisis va a utilizarlas varias veces aprovechando los resultados de una para alimentar a la otra.

La visualización es una herramienta fundamental. Una buena visualización te mostrará el patrón de los datos, cosas que tal vez no esperabas o te hará surgir nuevas preguntas. También puede ayudarte a replantear tus preguntas o darte cuenta si necesitas recolectar datos diferentes.

Los modelos son herramientas complementarias a la visualización. Una vez que tus preguntas son lo suficientemente precisas, puedes utilizar un modelo para responderlas. Los modelos son herramientas matemáticas o computacionales y tienen supuestos para poder aplicarlos, asi que la tarea de seleccionar el modelo adecuado para nuestro problema es una parte importante de este proceso, como también lo es su implemtación e interpretación posterior.

El último paso en el proceso de la ciencia de datos es la **comunicación**, una parte crítica de cualquier proyecto de análisis de datos, porque es cuando vas a mostrar tus resultados a otras personas y necesitas que puedan comprenderlos y encontrarlos útiles para utilizarlos.

Alrededor de todas estas herramientas se encuentra la **programación** como herramienta transversal en el proyecto de ciencia de datos. No necesitas ser una personas experta en programación para hacer ciencia de datos, pero aprender más sobre programar te ayudará a automatizar tareas recurrentes, compartir tu trabajo de forma reusable y aprovechar el trabajo de otros para resolver problemas similares con mayor facilidad y rapidez.

En este cuadernillo te mostraremos como realizar cada una de estas etapas utilizando el software R y te dejaremos links donde puedes aprender y profundizar más cada aspecto de este proceso.

1.1. ¿Por qué R?

Excel es un software admirable. Es genial para hacer data entry, para ver los datos crudos y para hacer gráficos rápidos. Si venís usándolo hace tiempo, seguro que aprendiste un montón de trucos para sacarle el jugo al máximo, habrás aprendido a usar fórmulas, tablas dinámicas, e incluso macros. Pero seguro que también sufriste sus limitaciones.

En una hoja de Excel no hay un límite claro entre datos y análisis. Sobrescribir datos es un peligro muy real y análisis complicados son imposibles de entender, especialmente si abrís una hoja de cálculo armada por otra persona (que quizás es tu vos del pasado). Además, repetir el análisis en datos distintos o cambiando algún parámetro se puede volver muy engorroso.

Si lo que necesitás son reportes frecuentes y automáticos, y análisis de datos con muchas partes móviles, estaría bueno poder escribir una receta paso a paso y que la computadora corra todo automáticamente cada vez que se lo pedís. Para poder hacer eso, ese paso a paso tiene que estar escrito en un lenguaje que la computadora pueda entender, ese lenguaje es R.

La forma en la que interactuás con la computadora con R es diametralmente distinta que con Excel. Esto lo hace extremadamente poderoso, pero el precio a pagar es básicamente el de tener que aprender un nuevo idioma.

1.2. Cómo decirle a R qué hacer

1.2.1. Orientándose en RStudio

En principio se podría escribir código de R con el Bloc de Notas y luego ejecutarlo, pero nosotros vamos a usar RStudio, que brinda una interfaz gráfica con un montón de herramientas extra para hacernos la vida más fácil.

Cuando abras RStudio te vas a encontrar con una ventana con cuatro paneles como esta:

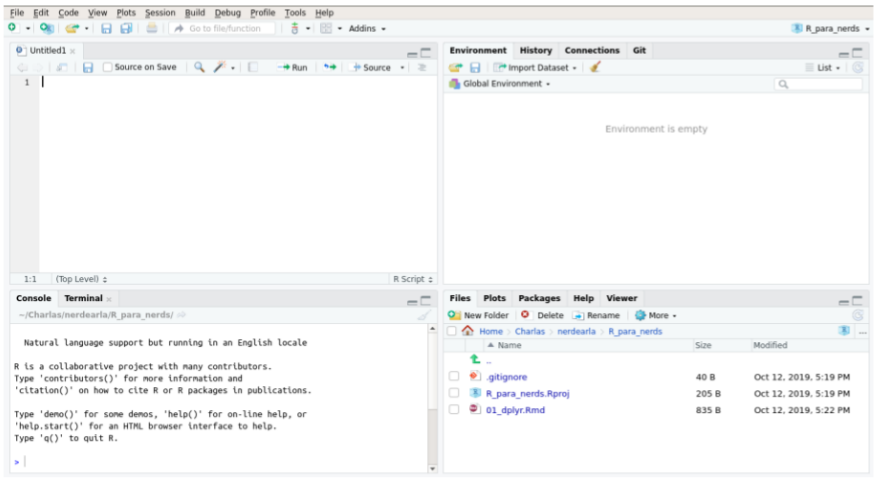

Figura 1.1: Ventana de RStudio

Los dos paneles de la izquierda son las dos formas principales de interactuar con R. El panel de abajo a la izquierda es **la consola**. Es el lugar que te permite *conversar* con R. Podés escribir comandos que se van a ejecutar inmediátamente cuando aprietes Enter y cuyo resultado se va a mostrar en la consola.

Por ejemplo, hacé click en la consola, escribí `2 + 2` y apretá Enter. Vas a ver algo como esto:

```
2 + 2
```

```
## [1] 4
```

Le dijiste a R que sume 2 y 2 y R te devolvió el resultado: 4 (no te preocupes del `[1]` por ahora). Eso está bueno si querés hacer una cuenta rápida o chequear algo pequeño, pero no sirve para hacer un análisis complejo y reproducible.

En el panel de arriba a la izquierda tenemos esencialmente un editor de texto. Ahí es donde vas a escribir si querés guardar instrucciones para ejecutarlas en otro momento y donde vas a estar el 87 % de tu tiempo usando R.

A la derecha hay paneles más bien informativos y que tienen varias solapas que vamos a ir descubriendo a su tiempo. Para destacar, arriba a la derecha está el "environment", que es forma de ver qué es lo que está "pensando" R en este momento. Ahí vas a poder ver un listado de los datos que están abiertos y otros objetos que están cargados en la memoria de R. Ahora está vacío porque todavía no cargaste ni creaste ningún dato. Abajo a la derecha tienen un explorador de archivos rudimentario y también el panel de ayuda, que es donde vas a pasar el otro 13 % del tiempo usando R.

Entonces, para resumir:

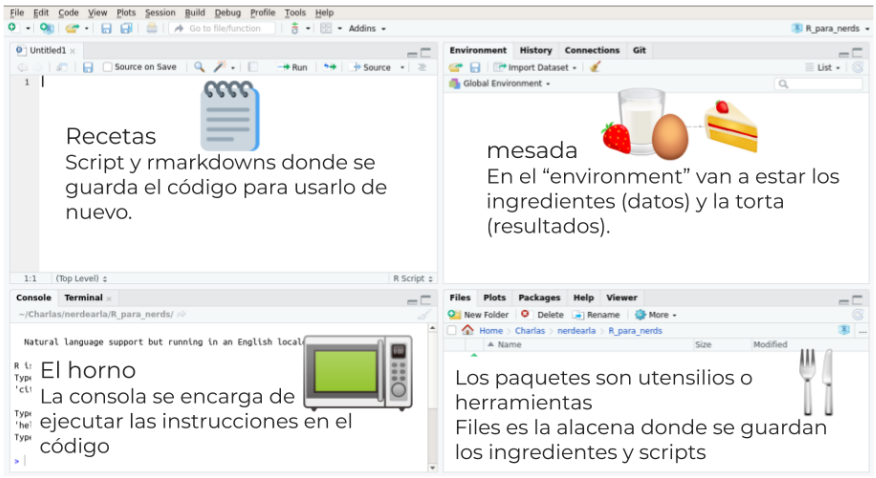

Figura 1.2: La cocina de RStudio

1.2.2. Hablando con R

Ya viste cómo usar R como una calculadora.

```
2 + 2
```

```
## [1] 4
```

Si usaste fórmulas en Excel, esto es muy parecido a poner `=2+2` en una celda. R entiende un montón de operaciones aritméticas escritas como seguramente ya te imaginás:

- `+`: sumar
- `-`: restar
- `*`: multiplicar
- `/`: dividir
- `^`: exponenciar

Pero además conoce muchas otras operaciones. Para decirle a R que calcule el seno de 1 hay que escribir esto:

```
sin(1)
```

```
## [1] 0.841471
```

Esto es similar a poner `=SIN(1)` en Excel. La sintaxis básica para aplicar cualquier función es `nombre_funcion(argumentos)`.

Nota: En Excel el nombre de las funciones dependen del idioma en el que está instalado. Si lo usás en español, la función seno es `SEN()`. En R, las funciones siempre se escriben igual (que coincide con el inglés).

Desafío

Decile a R que compute las siguientes operaciones:

- 2 multiplicado por 2
- 3 al cuadrado
- dos tercios
- 5 por 8 más 1

Al hacer todas estas operaciones, lo único que hiciste fue decirle a R que haga esos cálculos. R te devuelve el resultado, pero no lo guarda en ningún lado. Para decirle que guarde el resultado de una operación hay que decirle con qué "nombre" querés guardarlo. El siguiente código hace eso:

```
x <- 2 + 2
```

La "flechita" `<-` es el operador de asignación, que le dice a R que tome el resultado de la derecha y lo guarde en una variable con el nombre que está a la izquierda. Vas a ver que no te devele el resultado. Para verlo, ejecutamos