Data Capsule - Métodos Cuantitativos. Maestría en Gestión de la Innovación Social

Carlos Ignacio Patiño

Julio 2015

1. Introducción

El apropiado desarrollo y eventual implementación de métodos para el análisis de datos cuantitativos requiere siempre tener claridad acerca de la pregunta o hipótesis que se quiere responder o verificar. Contar con un rumbo o meta final permite agilizar el proceso previo al análisis de datos, llamado también análisis exploratorio. Este análisis exploratorio se centra en la aplicación de estadísitcas descriptivas (*summary statistics*) o medidas de tendencia central, y del uso extensivo de herramientas de visualización. En las primeras sesiones de este curso, cubriremos de manera muy práctica los conceptos de obtención, limpieza y análisis exploratorio de datos. Para tal fin, el presente ejercicio grupal y práctico, busca familiarizar a los estudiantes con el proceso de búsqueda de información, procesamiento y limpieza, y análisis exploratorio. Dado que se trata de un ejercicio de motivación, el estudiante está en la libertad de escoger el conjunto de datos que más le interese, o el que le sirva de fuente para responder a cualquier pregunta que tenga interés en responder. Sin embargo, por facilidad, en el presente ejercicio se sugieren algunos conjuntos de datos de interés, que pueden ser empleados por los grupos de trabajo para la realización de las actividades del *Data Capsule*.

El objetivo fundamental del presente ejercicio es obtener un mini reporte de resultados donde se describan, de manera breve, las conclusiones obtenidas durante el ejercicio exploratorio y donde se presenten también sugerencias de caracter de política pública o modelos de negocio en innovación social que permitan solucionar problemas detectados a partir del análisis de la información.

2. Instrucciones

2.1. Descargue y limpieza del conjunto seleccionado

Descargue y procese el archivo de datos seleccionado para su análisis exploratorio. En esta etapa, los grupos de trabajo pueden escoger su propio conjunto de datos de interés, o pueden seleccionar uno de los datos sugeridos para este ejercicio:

- Proyectos priorizados y viabilizados de agua (Ministerio de Vivienda): Proyectos priorizados y viabilizados en el marco del Programa Aguas para la Prosperidad SIGEVAS que le permite realizar la formulación, reformulación y evaluación de proyectos del sector de Agua y Saneamiento Básico. Disponible en: http://datos.gov.co/frm/catalogo/frmCatalogo.aspx?dsId=41015. Para descargar, dirigirse al link y dar click en cualquiera de los botones (Excel, o .csv).
- Encuesta de Uso del Tiempo (DANE): (Descripción tomada del DANE) Establecer cómo distribuyen el tiempo las personas en sus actividades cotidianas, ha sido una inquietud constante en investigadores alrededor del mundo. La ENUT recolecta información para la población civil no institucional residente en todo el territorio nacional, excluyendo los nuevos departamentos de la Orinoquia y Amazonia, el periodo de recolección de la información está comprendido entre el mes de agosto de 2012 y el mes de julio de 2013. La investigación está definida para tomar información de viviendas, hogares y personas, dentro de una estructura típica de encuesta de hogares. De las viviendas se indaga sobre: el tipo, la cobertura de servicios públicos y los materiales de los pisos. En cuanto a los hogares se pregunta por: la tenencia, la recepción de subsidios, la eliminación y separación de basuras, la tenencia y uso de bienes y recepción de ayudas en términos de trabajo por parte de otros hogares. Respecto

al apartado sobre las personas, la ENUT toma información demográfica básica y sobre salud de todas las personas del hogar, así como sobre el cuidado de menores de 5 años, la educación para las personas de 5 años, el mercado laboral y el uso del tiempo para las personas de 10 años y más. La documentación de todas las tablas incluidas en la ENUT se encuentra disponible en el siguiente link: http://formularios.dane.gov.co/Anda_4_1/index.php/catalog/214/data_dictionary. Los grupos de trabajo interesados en esta base de datos podrán acceder a ella a través del instructor, quien cuenta con una base de datos resumida (columnas seleccionadas) e integrada, para la facilitación del presente ejercicio. Adicionalmente, a estos grupos se les facilitará documentación adicional que explica el proceso de integración de las tablas en la base de datos de la ENUT, al igual que transformaciones ya realizadas a diversas columnas. La base provista por el instructor, se enfoca en las columnas correspondientes a la tabla que contiene las respuestas a las preguntas sobre uso del tiempo en educación. Con esta información, los estudiantes podrán investigar sobre los diferentes comportamientos que muestran las personas en cuanto al tiempo dedicado al estudio fuera de la jornada escolar.

- Datos generales de EAPB (Entidades Administradoras de Planes de Beneficios) (Supersalud): Disponible en http://datos.gov.co/frm/catalogo/frmCatalogo.aspx?dsId=41128. Para descargar, dirigirse al link y dar click en cualquiera de los botones (Excel, o .csv). En la pestaña "Ficha Técnica" podrá encontrar información acerca de las columnas incluídas en este archivo.
- Estadísticas en Educación Superior (Ministerio de Educación Nacional): Disponible en http://datos.gov.co/frm/catalogo/frmCatalogo.aspx?dsId=20674. Para descargar, dirigirse al link y dar click en cualquiera de los botones (Excel, o .csv). En la pestaña "Ficha Técnica" podrá encontrar información acerca de las columnas incluídas en este archivo.
- Víctimas por departamento y hecho victimizante (Unidad para la Atención y Reparación Integral a las Víctimas): Número de personas afectadas por hecho victimizante y departamento. Disponible en http://datos.gov.co/frm/catalogo/frmCatalogo.aspx?dsId=10544. Para descargar, dirigirse al link y dar click en cualquiera de los botones (Excel, o .csv). En la pestaña "Ficha Técnica" podrá encontrar información acerca de las columnas incluídas en este archivo.
- Accidentes de transito ocurridos en el municipio de Hato Corozal Casanare durante el año 2013. Esta información se encuentra disponible en http://www.datos.gov.co/frm/catalogo/frmCatalogo.aspx?dsId=62290.

Posterior al descargue de la información, proceda a abrir el archivo (Excel, R, SPSS, u otro paquete de su interés) y a revisar que la información se encuentre en buen estado y que no haya columnas o filas faltantes o con errores. La facilidad para hacer esta exploración visual dependerá del tamaño del archivo, tanto en filas como en su número de columnas (o variables). Archivos demasiado grandes requerirán de otros métodos de análisis y exploración que se salen del alcance de este curso.

2.2. Revisión de metadatos

Uno de los pasos más críticos en el análisis de información es la revisión de la documentación (metadatos) que usualmente se provee junto al conjunto de datos. La revisión de esta pieza fundamental de información es clave para la definición (o redefinición) de la pregunta o hipótesis de investigación. Usualmente, una detallada revisión de esta documentación permite reducir considerablemente el trabajo en pasos subsecuentes. Al mismo tiempo, seleccionar y conocer de antemano las variables que serán empleadas en analisis posteriores permite reducir el tamaño de la base de datos (más relevante cuando se trata de bases de datos masivas con más de 500 variables) y facilitar la manipulación de ésta en cualquier paquete estadístico. En este paso, los grupos de trabajo deberán dedicar un buen tiempo a la revisión de la documentación referente a la base de datos seleccionada. Para esta etapa, los grupos cuentan con 30 minutos.

2.3. Definición de la pregunta de investigación: Hipótesis

A partir de la inspección visual inicial, de su interés particular y de la revisión de la documentación provista por la fuente de informacion, escriba, de forma clara y concisa, la hipótesis que intenta validar a partir de la

información. El tiempo para la preparación de esta etapa es de 15 minutos.

2.4. Análisis

Los equipos de trabajo cuentan con 45 minutos para llevar a cabo cualquier análisis exploratorio que consideren pertinente y que les ayude a dar luces sobre la validez de la hipótesis planteada o sobre la respuesta a la pregunta definida en la etapa previa.

2.5. Recomendaciones de política o emprendimiento social

Durante esta etapa, los equipos de trabajo deberán pensar en posibles acciones de política (o de innovación social y emprendimiento) que permitan solucionar el problema detectado en las etapas previas de este ejercicio. El tiempo disponible para plantear una solución preliminar es de 15 minutos.

2.6. Preparación de un breve reporte

Los equipos de trabajo deben preparar una corta presentación (no más de 5 diapositivas), donde expliquen:

- La base de datos seleccionada
- La pregunta de investigación
- El análisis llevado a cabo
- Conclusiones principales
- Propuesta de política o de innovación social
- ¿Qué problemas encontró en la base de datos analizada? Incluya un par de sugerencias para los administradores de la base de datos, que faciliten futuros accesos a la información.

Para esta etapa los estudiantes cuentan con 15 minutos

2.7. Presentaciones grupos

Los 3 a 4 grupos de trabajo presentarán al resto de la clase el reporte de investigación. Cada grupo tendrá 10 minutos para presentar sus resultados.

3. Recomendaciones

- Este ejercicio será calificado y contará como parte de la nota final de participación y talleres en clase. Sin embargo, la calificación no se dará con base en la calidad y complejidad del análisis llevado a cabo, ni de las propuestas que surjan a partir de dicho análisis. La calificación se centrará en la calidad del trabajo en equipo y en la profundidad que se tenga en cuanto a la definición de la pregunta o hipótesis de investigación. Por lo tanto, se sugiere a los equipos de trabajo dedicar esfuerzos a la revisión de los metadatos, a la selección de las variables adecuadas, y a la apropiada definición y argumentación de la pregunta de investigación.
- El instructor estará disponible durante todo el ejercicio para resolver dudas y ayudar a los equipos en la descarga y carga de la información.
- La base de datos del DANE cuenta con un grado mayor de dificultad, debido a su tamaño y complejidad. Por lo tanto, equipos que decidan seleccionar dicha base para este ejercicio serán recompensados con puntos adicionales de bonificación, sujetos a la calidad del ejercicio inicial (exploración y definición de la pregunta).

- Equipos que seleccionen otras bases de información de su interés, diferentes a las sugeridas en este ejercicio, y que completen satisfactoriamente las etapas del ejercicio, también serán recompensados con puntos adicionales en su nota de participación.
- La duración total de este ejercicio será de 3 horas.